Mémoire présenté le :

pour l'obtention du Diplôme Universitaire d'actuariat de l'ISFA et l'admission à l'Institut des Actuaires

Par : DO Xuan Quang

Titre Predictive Underwriting vs Traditional Underwriting with Data Science approach

Confidentialité : ☐ NON ■ OUI (Durée : ☐ 1 an ■ 2 ans)

Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membre présents du jury de l'Institut des Actuaires signature

Membres présents du jury de l'ISFA

Secrétariat

Bibliothèque :

Entreprise :

Nom : AXA Global Life

Signature :

Directeur de mémoire en entreprise :

Nom : Sami Faye-Chellali

Signature :

Invité :

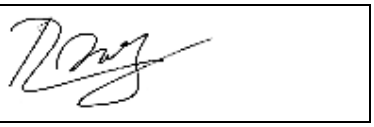
Nom :

Signature :

Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels *(après expiration de l'éventuel délai de confidentialité)*

Signature du responsable entreprise

Signature du candidat



DO Xuan Quang

Tutor: Sami Faye-Chellali

AXA Global Life

Actuarial thesis
Master of Actuarial and Financial Science (SAF)

Predictive Underwriting vs Traditional Underwriting with Data Science approach

April 5, 2017

Abstract

Predictive modeling is poised to alter how the industry rates and underwrites [1]

Nowadays, insurers face many intense competitions on all the value chains of the business, not only on the pricing but also on the risk acceptance. The Individual Protection sector is not out of the competition. To become and remain as a competitive actor on the market, insurers must attract more clients, constructing less risky insured portfolio associated with a profitable price. One of the possibilities is to propose the less complex underwriting process, and eventually more segmentation. Indeed, on such a complicated guarantee like protection, the faster the underwriting process is, the more clients we have in our portfolio. Because with the same level of guarantees, a particular client will move more readily towards the insurer that he/she can buy the product within only one interview (ideally on the internet) rather than the insurer requesting additional laboratory analysis, even if the price is a little bit higher.

With this motivation, under the development of Data Science framework, insurers have two alternatives. The first one is to adapt the traditional underwriting process in order to have the better risk assessment with simplified underwriting process. This alternative seems challenging, and still, the cost paid to the underwriters remains important. The second alternative is to develop mathematical algorithms called predictive underwriting, an intelligent way to do the underwriting. It is not just a tool to do more business, but also a tool for the risk control. The advantage of the predictive underwriting is that it can produce more or less the same decision as the traditional underwriting with less underwriters, fewer questions, hence less complicated underwriting process, and more customer friendly. But the consequence is that the predictive underwriting at the first stage still has some (even little) prediction errors. That is to say, we accept the clients associated (by the traditional underwriting) with high level of risk that we shouldn't take into our portfolio. To mitigate this type of risk, we have to apply some extra loading on the premium.

How to develop the predictive underwriting model? How much is the prediction errors? How can the price be adjusted in order to capture the new risk entering into the portfolio? This thesis will answer this 3 questions in detail, not only on the technical perspective, but also on the business point of view.

The structure of this thesis is divided into three main parts:

1. Context, motivation of the study: talk about the need of underwriting in protection business, and explain how can Data Science be involved to improve this process.
2. Machine Learning, Predictive Underwriting: develop in details two models for predictive underwriting: the first one is *Logistic Regression*; and the second one is *Random Forest*. After building the model, we discuss about their potential applications in different value chains of the business such as: reduction of the underwriting circle, improvement of telemarketing approach, etc.
3. What are the consequences of Predictive Underwriting? How does the structure of our insured portfolio change? How much should we adjust the price in order to mitigate the risk of portfolio deformation?

Key words: Predictive Underwriting, Machine Learning, Data Science, pricing, random forest, logistic regression, mortality table, protection, generalized linear model.

[1] Insurance-canada.ca

Résumé

Dans un monde moderne, les assureurs font face à de nombreuses compétitions intenses sur toutes les chaînes de valeur de l'entreprise, non seulement sur le prix mais aussi sur l'acceptation du risque. Le secteur de la prévoyance individuelle n'est pas hors compétition. Pour devenir et maintenir un acteur concurrentiel sur le marché, les assureurs doivent attirer plus de clients moins risqué avec un prix rentable. L'une des possibilités est de proposer le processus de souscription moins complexe et éventuellement une segmentation de risque plus profonde. En effet, sur une garantie aussi complexe que la prévoyance individuelle, plus le processus de souscription est rapide, plus nous avons de clients dans notre portefeuille. Parce qu'avec le même niveau de garanties, un client préfère l'assureur avec lequel il peut acheter le produit en une seule entretien (idéalement sur Internet) plutôt que l'assureur qui demande une analyse de sang supplémentaire, même si le prix est un peu plus cher.

Avec ce motif, et dans le cadre du développement de la Data Science, les assureurs ont deux alternatives. Le premier consiste à adapter le processus de souscription traditionnel afin d'avoir une meilleure évaluation des risques grâce à un processus simplifié de souscription. Cette alternative semble difficile, et encore, le coût payé pour les souscripteurs reste important. La deuxième alternative est de développer des algorithmes mathématiques appelés souscription prédictive, une manière intelligente de faire la souscription. Ce n'est pas seulement un outil pour faire plus de business, mais aussi un outil de contrôle des risques. L'avantage de la souscription prédictive est qu'elle peut produire plus ou moins la même décision que la souscription traditionnelle avec moins de souscripteurs, moins de questions, donc un processus de souscription moins compliqué et plus convivial pour les clients. Mais la conséquence est que la souscription prédictive à la première étape présente des erreurs de prédiction. C'est-à-dire que nous acceptons les clients associés (par la souscription traditionnelle) à un risque élevé que nous ne devrions pas prendre dans notre portefeuille. Pour atténuer ce type de risque, nous devons appliquer un chargement supplémentaire sur la prime.

Comment développer le modèle de souscription prédictive? Combien coûtent les erreurs de prédiction? Comment le prix peut-il être ajusté pour mutualiser le nouveau risque entrant dans le portefeuille? Ce mémoire répondra à ces 3 questions en détail, non seulement sur la partie technique, mais aussi sur le point de vue commercial.

La structure de ce mémoire contient trois parties principales:

1. Contexte, motivation de l'étude: discuter du besoin de la souscription dans le business de la prévoyance individuelle et expliquer comment la Data Science peut être appliquée pour améliorer ce processus.

2. Machine Learning, souscription prédictive: développer en détail deux modèles de souscription prédictive: *Logistic Regression* et *Random Forest*. Après avoir construit le modèle, nous discutons de leurs applications potentielles dans les différentes chaînes de valeur de l'entreprise telles que: la réduction du cercle de souscription, l'amélioration de l'approche de télémarketing, ...

3. Quelles sont les conséquences de la souscription prédictive? Comment la structure de notre portefeuille assuré change-t-elle? Combien devrions-nous ajuster le prix afin de mutualiser le nouveau risque de déformation du portefeuille?

ACKNOWLEDGMENTS

It has never been easy for a student to find an internship, so I want to thank AXA Global Life for welcoming me during these six months and for the efficient collaboration.

First of all, I express my gratitude to Sami Faye-Chellali, Global Head of Indivudual Protection Strategy & Offer and my supervisor at AXA Global Life, for giving me a chance to work on an extremely interesting subject, for the enormously valuable guidance all along my internship, for many constructive and encouraging work sessions, for the appreciation on my work, and for the authority to present my work to many AXA's entities.

I want also to thank Sabrina Ledent and Caroline Calov, my actuarial colleagues for giving me many clear explanations for my daily questions concerning different actuarial issues in the professional environment.

My thanks go also to Enrique Lopez at AXA Global Life and Mustapha Lakehal at AXA Group Solutions for their excellent collaboration helping me to quickly understand the exigence of the project and different details on the data.

I would like to thank also members of the Underwriting team, including Sohel Abu (group chief medical officer), Henriette Carvalho and Tim Romain (senior underwriters) for their help in order to better understand the different phases of underwriting process implemented at AXA.

Finally, I want to thank members of AXA Innovation Lab and AXA MPS (Italy) for their technical support and for providing internal as well as external data used in the model.

Contents

I Context **6**

1 Presentation of the study **7**
1.1 Presentation of AXA . . . 7
1.2 AXA Global Life . . . 7
1.3 Individual Protection team . . . 8
1.3.1 What is Protection? . . . 8
1.3.2 Why should Protection be a priority of AXA? . . . 9

2 Goal of the study **11**
2.1 Traditional Underwriting . . . 11
2.2 Underwriting process in protection . . . 12
2.3 Innovation: Automated Underwriting . . . 16
2.4 Ambition: Predictive Underwriting . . . 18
2.5 Comparison of Traditional Underwriting and Predictive Underwriting . . . 19

3 Data presentation and descriptive statistics **21**
3.1 Type of data used in Predictive Modeling . . . 21
3.2 Some descriptive statistics on AURA's data . . . 22
3.3 External data . . . 31

II Machine Learning, Predictive Underwriting and Client Scoring **32**

4 Logistic Regression **34**
4.1 Introduction . . . 34
4.2 Generalized Linear Model . . . 35
4.3 Dichotomous logistic regression . . . 36
4.3.1 Introduction and fundamental hypothesis of the logistic regression . . . 36
4.3.2 Fitting the model by the maximum likelihood estimation . . . 38
4.3.3 Estimation of the standard errors . . . 39
4.3.4 Significance test of the coefficients . . . 40
4.3.5 Confidence interval of the estimation . . . 43
4.4 Polychotomous logistic regression . . . 43
4.4.1 Fitting the model . . . 43
4.4.2 Significance test . . . 45
4.5 Model building strategies . . . 46
4.5.1 Introduction . . . 46

4.5.2 Variable selection . . . 46
4.5.3 Selection of the preliminary model . . . 49
4.5.4 Selection of the final model . . . 50
4.6 Evaluation of the regression model . . . 54
4.6.1 Goodness of fit . . . 54
4.6.2 Evaluating the prediction power - the validation set approach . . . 62
4.7 Application (in underwriting and client scoring) to the AXA Italy portfolio . . . 63

5 Random Forest **66**
5.1 Tree-based methods . . . 66
5.2 Bagging and Random Forest . . . 69
5.2.1 The bootstrap aggregation or bagging . . . 69
5.2.2 Random Forest . . . 70
5.3 Comparison of the results obtained from different methods . . . 86

6 Impact of Predictive Modeling **87**
6.1 Pricing of the life insurance policies . . . 87
6.2 Pricing taking into account the prediction errors . . . 90

7 Conclusion and Further Room for development **95**

Appendices **99**

A Result of the `glm` function in R **100**

Part I

Context

Chapter 1

Presentation of the study

1.1 Presentation of AXA

The AXA Group, world leader in Financial Protection, supports and advises its individual and corporate customers at every life stage, providing them with the products and services that meet their insurance, personal protection, savings and wealth management needs.

Our areas of expertise are reflected in a range of products and services adapted to the needs of each client in three major business lines: property-casualty insurance, life & savings, and asset management. Present in 57 countries, the 163 000 employees and distributors of AXA are committed to serving 101 million clients.

1.2 AXA Global Life

AXA Global Life is a transversal organization that aims to develop synergies within the Group in the life insurance sector with a view to increasing our revenues, profitability and efficiency.

The structure of AXA Global Life is divided into two mains parts corresponding to its responsibilities. There is one part acting as an internal reinsurer of all AXA's entities for life business. Up to now, AXA Global Life's reinsurance manages 71 life reinsurance contracts corresponding to €11,3 millions of premiums ceded. The second structure acts as a center of global managers who define scopes, long-term and short-term strategies for all AXA's entities for life, savings & health business of the group, in which the premiums under the management accounts for billions of euros.

The GBL is responsible for the global steering of the business in line with the Group rules, standards and policies.

Our mission & activities

Our strategic priorities are to:

- Diversify the personal protection and health activities;
- Reestablish profitability in the savings sector;

- Target investments in order to capitalize on opportunities for growth;
- Improve the profitability of our business in order to continue financing our growth.

These priorities have been defined in a restrictive economic environment, characterized by strong pressures over the Group's growth and profitability.

They are being rolled out through 6 main initiatives, encouraging collaboration with our partners in the companies, with a view to developing our international business and achieving our goals together.

AXA's Global Program for L&S

Key figures

- 20 nationalities and several languages spoken (French, German, Dutch, Spanish, Italian, Vietnamese, Korean, Moroccan, Chinese, British English...)
- Around 70 employees in total
- 25 employees organized in 5 competence centers and some support functions (CFO, Head HR, Assistant)
- Reinsurance is a separate organization
- 57 life entities we are working with

1.3 Individual Protection team

1.3.1 What is Protection?

In insurance, Protection business refers to products providing coverage for persons, in opposition to Property & Casualty insurance, or Savings products. It comes in the form of a money benefit (sometimes packaged with additional services) provided to the beneficiary after the occurrence of a specified event. The triggering event can be death, invalidity, illness, dependency, accident, hospitalization, etc ...

The definition encompasses a large variety of products, which can be classified in different categories within the different entities of AXA, following historical or strategic reasons, or market practices.

Protection products are more generally classified in the Life & Savings segment under the two following categories:

- Pure Protection products.
- Protection with Savings components products.

In practice, the frontiers between Protection business and other insurance segments are today not clearly drawn due to three different unclear boundaries:

- Frontier between Protection and P&C (question of Personal Accident products)
- Frontier between Protection and Health (question of Long-Term Care, Hospitalization, Critical Illness, etc...)
- Frontier between Protection and Savings (question of Protection with Savings products)

The classification of Protection products has been guided so far by products features. As an example, personal accidents products are often classified as P&C as one year small premium products, not requiring a strong advice dimension, and often sold together with other P&C products; similarly, underwriting and claim management for these products are closer to P&C business.

1.3.2 Why should Protection be a priority of AXA?

Protection products are the core business of Life Insurance, as they meet the primary customer demand to cover unexpected events having a strong impact on their life. Even more than P&C products, Protection products address emotional issues such as death, disability, accident, illness, etc..., and consequently would play a central role in AXA's Customer Centric value proposition.

Protection products still remain very attractive for financial reasons. In particular, Individual Pure Protection offers a very high profitability. The IRR for Protection & Health was higher in 2013 compared to in L&S.

In addition, the Protection business shows a huge growth potential that can be illustrated with the current Protection gap (defined for individuals as the difference between resources and the needed amount to maintain their living standards). The mortality Protection gap was estimated to represent €64 trillion worldwide in 2012 (compared to €2.6 trillion global Life insurance premiums), leaving a tremendous growth potential for Protection insurance in the future years. The growth of Protection business should accelerate in the coming years due to a combination of social, economic and demographic factors: the willingness to buy Protection coverage is notable positively correlated to wealth and age.

- Protection coverage demand is increasing with wealth: In the coming years, the middle class population is expected to grow exponentially, as emerging countries will significantly increase their wealth. The OECD forecasts an exponential growth of the global middle class population: it would be multiplied by 2.6 between 2010 and 2030, from 1.8 to nearly 5 billion inhabitants. This increase is expected to be largely beyond the global population expansion, since the middle class would represent more than the half of the global population, whereas it covered only 27% of it in 2010.

- Protection coverage is increasing with age: The global population is drastically ageing: the 65+ population is expected to increase at a 2.8% CAGR between 2015 and 2030, from 600 million in 2015 to more than 900 million inhabitants in 2030 , when the total population is only expected to grow at a 1.0% CAGR. The share of 65+ inhabitants will grow from 8% in 2015 to 11% in 2030. In this context, life expectancy is growing with a pronounced pace, by more than 2 years over this period. Consequently, the growing "silver economy" market should affect deeply the demand for Protection products from senior people over the period covered by the Vision 2030. This demand could provide a sustainable growth in a large scope of products such as whole life, long-term care, disability coverage...

As of today, the Protection market can be estimated on a conservative basis to represent around €400 billion premiums worldwide (15% of the total Life & Savings insurance market), although it is difficult to size precisely due to the lack of consistence in the scope from one country to another. This figure only accounts for premiums dedicated to coverage of risk and therefore excludes any Savings component. As highlighted above, the absence of significant transformation of the business identified in the coming years coupled with the fact that no new competition from non-insurance players is expected should be considered as a clear opportunity for AXA to position itself as a leader in Protection. While there is a significant risk that insurers' margins shrink in the future for Motor or Health products due to the pressure of new entrants across the value chain, the profit pool of Protection will fall in majority into insurers' pockets.

Chapter 2

Goal of the study

One of the important missions of the team *Individual Protection strategy and offer* is to create new models that can be applied in order to improve the performance of the business. An initiative is to apply Big Data / Data Science to improve the underwriting process. This thesis will answer five questions:

1. What is underwriting?
2. Why do insurers need to underwrite?
3. How is underwriting done in the traditional way?
4. What is Big Data?
5. How could Big Data affect the underwriting process and bring business value?

2.1 Traditional Underwriting

Protection is a special type of insurance because policyholders can have more information on (or at least different interpretation of) their health situation than insurers. In this case, we talk about the issue of information asymmetry. As the consequence, if insurers don't make the good segmentation or "imbalanced" pricing they will suffer from two kinds of risk:

1. Moral hazard
2. Anti-selection

The economical approach on the asymmetry information was principally initiated in economics of insurance, by two founder articles : one of Arrow (1963) and the other of Rothschild and Stiglitz (1967).

Arrow mentioned principally about the "adverse selection of risk" that can be generated by the imperfect pricing, because of the fact that the insurers who don't estimate sufficiently good their insured risks would be attractive for the bad risks on the market.

Moral hazard

Moral hazard concerns the influence of the insurance contract on the behavior of the insured. In fact, when an economic agent is assured, its behavior of "risk management" will be influenced by insurance coverage, which can generate an increase amount of risk. In the case of protection products, an individual may, for example, become less vigilant about their health situation because of a "don't worry, it's insured". By offering payouts to protect against losses from incidents, insurers may actually encourage risk-taking, which results in them paying more in claims.

Moral hazard can not be detected at the underwriting phase but can be reduced through the integration of some elements in the contract:

- *Waiting period*: this refers to a period fixed by the contract during which incidents occurring are not claimable. Adding a waiting period allows to avoid affections or incidents occurring in a concomitant way just after underwriting the contract and to better preserve the specific hazard of any insurance transaction.
- *Guarantee exclusions*: it is a clause that define the extent of the policy by cutting down the scope of cover. The cutting could be:
  - *Direct*: excluding the insurer from liability for a particular risk.
  - *Indirect*: delimiting (describing) the risk.

Anti-selection

For example, we have 2 insurers named A and B providing auto-insurance on the market. They propose the contract covering exactly the same accident risk. The only difference between 2 insurers is their price segmentation.

- Insurer A proposes 100 euros for 1 year insurance to all type of prospects.
- Insurer B is able to distinguish bad drivers and good drivers thanks to their underwriting process. Hence, they propose the price of 80 euros for good drivers, and 120 for bad drivers (Suppose that the proportion of good drivers is equal to that of bad drivers).

So, at the first look, one can say that insurer A and insurer B gain the same amount of profit ($\frac{80+20}{2} = 100$). But, from the insured point of view, if one is classified as a good driver, facing two different proposed prices, he/she will surely chose the contract with better price, i.e insurer B. Same thing for bad drivers, they will go to insurer A. As the consequences, the portfolio of insurer A has plenty of bad drivers, with the price calculated on the mixed portfolio, i.e a portfolio composes of bad drivers and good drivers, and vice versa for the insurer B. Finally, insurer A ends up by paying more claims than what it received in term of premium.

To be more competitive on the market, the challenge of the insurers is to minimize the anti-selection effect.

2.2 Underwriting process in protection

The function of the insurers is to cover the financial impact deriving from the unexpected occurrence of the events that negatively affect the insured. Insurance companies offer the claim payments in

exchange for a determined amount of premium. Based on the central limit theorem, insurers pull the similar risks together in order to transform the unpredictability into expected events happening to anyone in the pool. One of the assumptions of central limit theorem is that the risks must be homogeneous. Thus to apply correctly the central limit theorem, insurers must reduce the volatility of the portfolio by creating the different sub-portfolios of homogeneous risk with large enough number of insured. In reality, insurers accept the insured having different risk exposures but with the different conditions such as: applying debits (additional prime), excluding some types of risks covered, or waiting periods to restore the necessary risk homogeneity in the portfolio.

Underwriting refers to the process of risk selection and classification. Depending on the information and evidences provided, the customer's risk is carefully evaluated and appropriate premium is determined. Provided all information were in good faith and the policy will continuously be in force, this set premium is deemed binding to the insurance company.

Underwriter - The term "Underwriters" refer to professionals who perform the task of risk selection and classification based on a set of guidelines.

Risk selection is the method employed by underwriters to decide whether the risk is insurable, uninsurable or insurable with modified terms (exclusion, reduced duration). On the other hand, risk classification determines the insurance premium that should be charged based on the risks being presented by the customer (standard, substandard).

There are 2 types of risks:

1. **Speculative or Dynamic Risks** - refer to risks wherein profit or loss is feasible. For example, a person getting a life insurance on his/her mentally ill neighbor. In this scenario, the person purchasing the insurance cover will gain a lot upon the neighbor's demise. This type of risk should NOT be insurable and we should guard our company from these types of risks.

2. **Pure or Static Risks** - are risks that we want to cover. This refers to situations with only possibilities of loss or no loss and with no chance of gain. Examples are premature death, unexpected medical expenses, acute or chronic disability, etc.

There are 5 major risk types:

1. **Preferred Class** - a risk class of healthy prospects who are in excellent health, have a good family history and demonstrate favorable behaviors. The anticipated mortality is significantly lower than average. The preferred risk class may be broken down into different levels whose terminology may vary by product and market. The best preferred rates are usually granted to the healthiest individuals with the most favorable risk factors. Premium savings can total as much as 30% when compared to the standard premium. However, this risk class also requires many additional pieces of evidence such as medical records, medical tests, motor vehicle reports and others to establish the customer's mortality risk.

2. **Standard Class** - a risk class of proposed insured, whose anticipated mortality/morbidity risk is within that of an average, insured individual. Applications are approved and standard premiums rates as determined by the pricing team are applied.

3. **Substandard Class** - risk classes of proposed insured, whose anticipated mortality/morbidity risks are higher than that of average insured individuals, but are still considered insurable.

Applications are approved with modification on the premium rates or plan/riders or extent of coverage. Provided that the risk is assessed and priced accurately, we believe that substandard class is a profitable business.

4. **Postpone** - a risk class of proposed insured whose anticipated mortality/morbidity risk cannot be quantified during application usually due to limited information or lack of disease control. Examples include uncontrolled hypertension, diabetes, previously undisclosed growth during medical exam, abnormal stress test, etc.

5. **Decline** - a risk class of proposed insured whose anticipated mortality/morbidity is so great that we either cannot provide coverage at an affordable cost or cannot assess the risk.

Some fundamental variables that the underwriters usually take into account are: personal information (age, gender), current physical condition, personal medical history, habits (such as smoking, alcohol drinking, ...), family medical history, occupation, financial situation, avocation, residence and travel, beneficiary designation, ...

Factors considered in Risk Selection?

1. **Personal Information**

   - **Age** - except for the first few years of life, an individual's resistance to diseases and injuries goes down with the passage of time resulting to higher morbidity and mortality with increasing age.
   - **Gender** - in some markets (Europe), gender is no longer considered as a factor in risk selection as it is deemed discriminatory. However, in markets where gender premium rate differentiation is allowed, such is anchored on studies that show higher mortality for men and higher morbidity in women.

2. **Current Physical Condition** - This refers to body mass index (BMI), blood pressure and presence of developing illnesses that may be noted during routine medical tests as may be required by AXA. This information should be well taken into consideration as the extent of risk that we take on for certain lives should be appropriately priced.

3. **Personal Medical History** - This refers to customer's record of adverse medical conditions, injuries or operations which may develop into complications thereby affecting the individual's overall well-being in the future. Information on date of diagnosis, treatment (commencement and end date), current medication, and medical follow-ups should be carefully scrutinized and considered in the overall underwriting decision.

4. **Habits** - This refers to smoking (cigarettes, e-cigarettes and tobacco), alcohol drinking and use of prohibited drugs. AXA entities offer smoker and non-smoker differentiated insurance plans and premium rates for smokers are expectedly higher in view of the fact that such habit can aggravate a number of health conditions. Sometimes the smoker premium rates may not be enough to fully cover the extra risk particularly for heavy smokers. As such, it is prudent to always check on the number of cigarettes/tobacco consumed on a daily basis and the number of years that the customer has been a smoker. The appropriate substandard rate on smoking habits should then be added to other co-existing medical conditions such as diabetes, coronary artery disease, etc. (if any). Customers with excessive alcohol drinking may be accepted at substandard rates albeit depending on results of liver function tests and presence of any other medical conditions. Worst case scenario, the case may have to be declined. On cases where

there is a declaration of use of illegal drugs, the decision is typically to decline the application. For customers who had recreational or occasional use and who had not used it for an extended period of time (at least 2 years), we may consider acceptance of the application taking into account the customer's occupation, reliability, etc.

5. **Family Medical History** - Family medical history is another factor that should be considered in underwriting in view of certain medical conditions that are transmitted through the genes and therefore increasing a person's propensity to develop such disease in the future. This should strongly be considered when underwriting critical illness or where there are more than 2 family members who suffered from the disease before age 60.

6. **Occupation** - Occupational risks used to be much more significant than today. But with the implementation of safety measures in numerous industries, occupational risk has decreased. However, this is still a continuing concern for occupations that expose individuals to inhalation of harmful chemicals or increased risk for accidents and death.

7. **Finances** - The amount of insurance being applied for has to be weighed against the customer's income and assets. As we should only deal with pure risks, we should make sure that the amount of insurance being applied for is justified and there should be no gain at the time of claim. In order to determine the justification of the insured amount, the AXA Group financial UW guidelines set forth on Section 8 of this document must be followed. Where within Group insurance individual underwriting rule applies, it is expected that all entities adhere to these financial UW guidelines.

8. **Avocation** - This refers to an individual's hobby. Focus should be made on avocations that may present increased risk for accident or even death such as scuba/sky diving, hang gliding, rock climbing, competitive racing, etc. Oftentimes, flat extra rating has to be imposed or at worst, the case will have to be declined.

9. **Residence and Travel** - Residence or prolonged travel to foreign countries may at times require imposition of additional flat extra premium. Factors such as access to medical facility, socio-political situation, endemic diseases, access to reliable claims documents, etc. should be carefully considered.

10. **Beneficiary Designation** - As a general rule, designated beneficiary/ies should have insurable interest on the continued life of the proposed insured. Insurable interest exists when the designated beneficiary has more to gain when the insured continues to live than when the insured dies. Although this is not strictly being implemented nowadays, it is appropriate that guidelines are created around this to mitigate risk for early claims and possible insurance fraud.

Where to get the information?

The insurance companies collect information through many sources such as:

1. **Application Form**: The application form is the primary and legal basis of the insurance contract. An accurately and properly completed application form should provide the underwriter a complete picture of the customer.

2. **Physical Examination/Full Medical Examination (FME)**: This should be performed by an AXA accredited doctor or paramedic and the results of the examination should be recorded in an AXA prescribed form which should be signed by the customer and the examiner. Part 1

of the examination is focused on gathering the customer's personal and family medical history and Part 2 would require actual physical examination such as blood pressure, height and weight measurement, auscultation of the lungs, heartbeat, palpation of the liver, breast examination, etc.

3. **Medical Tests**: This refers to routine urinalysis, blood tests, ECG or Treadmill Stress Test. It is an accepted fact that some medical conditions (known or unknown to the customer) may not be observed during the physical examination. As such, blood tests, ECG or Treadmill Stress tests are required for high amounts of insurance cover, higher ages and for customers with known medical conditions. These medical test results should be properly evaluated.

4. **Financial Evidences**: Financial evidences or proofs of income such as validated tax returns, properties, and audited financial statements of the business may be required, depending on the plan and total financial sum at risk.

5. **Underwriting Questionnaires**: These are supplementary questionnaires which the customer has to complete in order to provide the underwriter more information about his or her medical condition, occupation, avocation and family history. With comprehensive information in the questionnaire, some medical tests may be waived and an underwriting decision may be reached thus saving on medical cost without compromising the quality of risk selection and overall cycle time is reduced.

6. **Attending Physician's Statement (APS)**: In some countries, APS may also refer to Medical Attendant's Report (MAR). Whenever necessary, this may be required for customers with adverse medical history. This should be considered mandatory for customers with heart, psychiatric, neurological conditions and history of cancer and for older customers who are getting high amounts of cover. For some markets like US, a comprehensive APS and some blood tests may be accepted in lieu of a routine physical examination or FME.

7. **Inspection Reports**: This refers to life inspection usually done by an independent third party. In some markets, direct interview of the customer is made and in others, this is done discreetly. Customer's income, assets, community reputation and medical conditions are checked. This report is typically being required on customers who are applying for high sum assured. However, with the advent of internet usage, electronic searches of public databases on property ownership, criminal and driving records, and credit history may be done even for moderate covers.

8. **Medical Information Bureau (MIB)**: This is available in some markets such as the US and other Asian countries and is basically a repository of confidential information. Most of the information is medical in nature, specifically on customers who had some critical information noted during underwriting which resulted to the application being accepted on modified terms, postponed or declined. Member insurance companies maintain and access this database. Information is coded and database is highly confidential and computerized. MIB was created to help protect insurance companies from insurance fraud.

2.3 Innovation: Automated Underwriting

Until now, most of the underwriting process is done manually, and is time consuming. Furthermore, underwriting information is difficult to be gathered even the basic information of the customer, and almost impossible to combine the customer information with the external information. In order to

reform and innovate the traditional approach, AXA decided to implement an automated tool developed by RGA (Reinsurance Group of America) named AURA (Automated Underwriting and Risk Analysis). AURA is a full decision tree which enables us to do real-time underwriting. It allows us not only to save time, to reduce the underwriting circle, so that we can sell more insurance contracts than with the traditional way, but also, it saves time for the underwriters so that they can focus on the more complicated cases.

Here is the big picture of how AURA works.

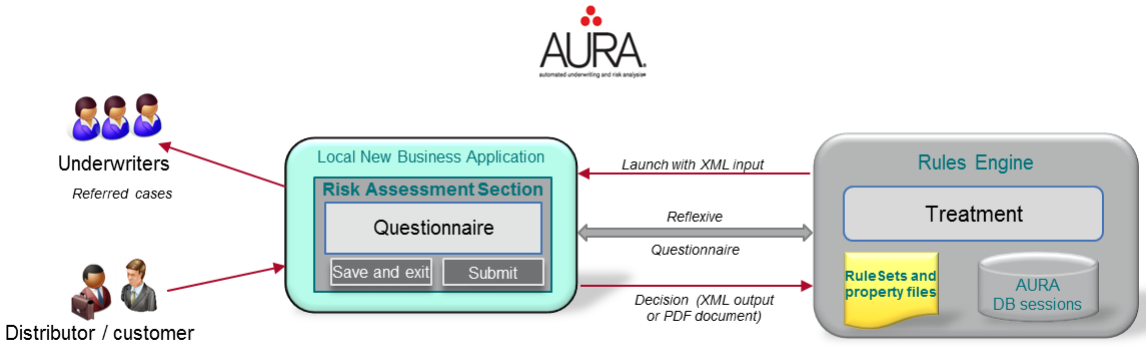


Figure 2.1: The structure of AURA

When a customer comes to our distributor to request an insurance quote, he/ she will work firstly with the distributor to answer the questions from the questionnaire powered by AURA. In the questionnaire, we have two kinds of questions: non-medical questions including: *Age*, *Gender*, BMI (we don't ask directly the BMI of customer because it makes things difficult to some people, instead we ask the substitute information such as $Height$ and $Weight$) , *Occupation*, *Sport Avocation*. The second kind of questions is the medical questions that allow us to have more details on customers' health. The answer of the customer in the questionnaire will go through the rule sets (decision tree) of AURA and AURA will make the real-time decision. The 4 decisions given by AURA are:

1. Accepting with standard, representing people with the good health conditions.
2. Accepting under some conditions (such as extra loadings), corresponding to people not having the good health conditions but still insurable. To mitigate this risk, we charge them some extra premium.
3. Referring to the underwriter, representing prospects that we are not sure about their insurability. These cases need further involvement of the underwriters.
4. Declining, corresponding to people having really bad health condition as we can not give them an insurance contract.

The accepted, accepted with extra loading, and referred cases are concluded directly by AURA. And the referred cases are sent back to the underwriters for further detailed analysis, for instance, blood testing, cholesterol testing, ... Unlike traditional underwriting in which there is no possibility to do statistics or analytics because the underwriting information is paper-based and is stored

physically in the data warehouse, if we need a specific information about an individual, we have to come to the data warehouse searching for the paper recording his/her information, the process is done physically. AURA allows us not only to store data in the cloud, but also to extract the data easily to do whatever kind of actuarial analysis. The inconvenience of AURA is that we just try to replicate (one part of) the brain of the underwriters in the computers, AURA ask the same number of questions to the client, and it remains less smart/accurate than the underwriters. The advantages of automated underwriting over traditional underwriting are:

1. It enables the reduction of the underwriting circle, hence we can make more business.
2. Cost reduction as it reduces the number of working underwriters.
3. AURA helps us to improve the underwriting performance, because underwriters can spend time concentrating on the more complex cases.

2.4 Ambition: Predictive Underwriting

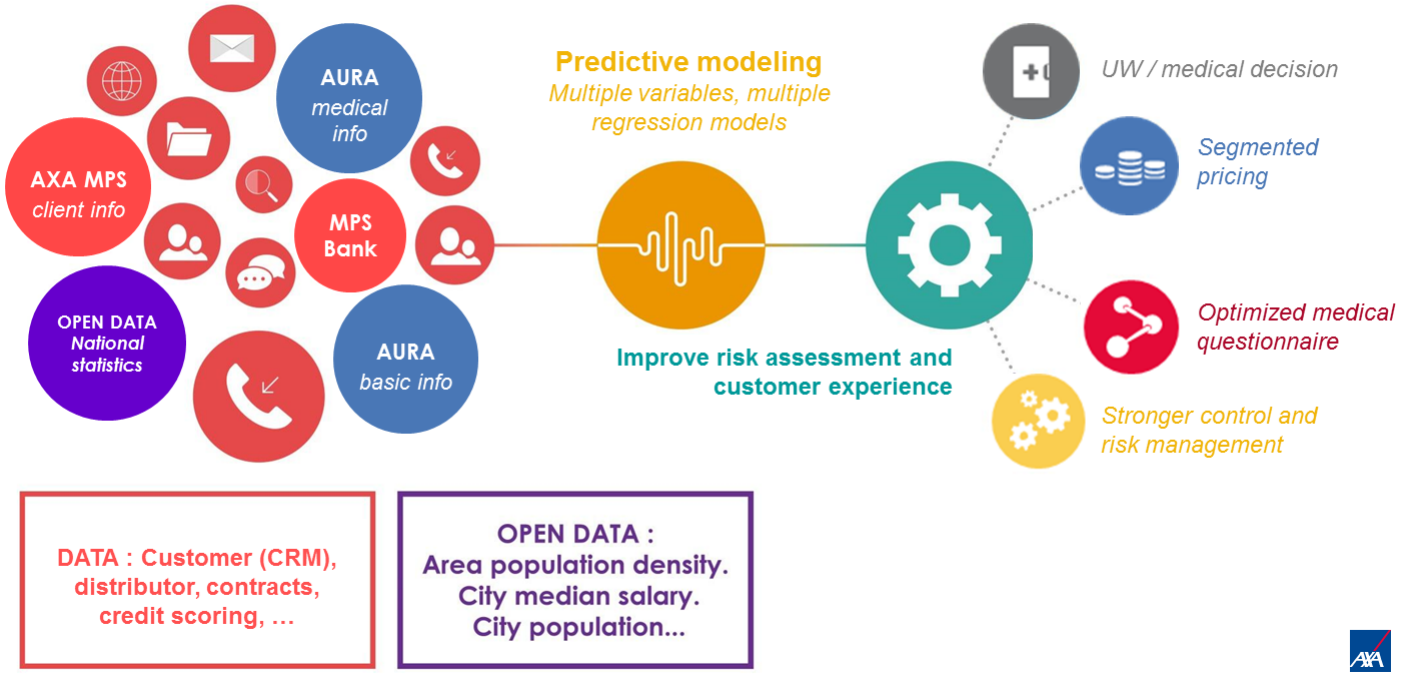


Figure 2.2: Predictive Underwriting

With the predictive underwriting, we try at the first step to deliver the same decision as AURA but with fewer questions in the questionnaire. Predictive Underwriting has various advantages comparing to traditional underwriting or automated underwriting. In the future, with the development of Big Data, we will be able to integrate the internal as well as external data to enrich the model taking even

fewer questions asked to the clients and delivering higher level of customer classification. Predictive Underwriting can:

1. Anticipate customer's profile, future behavior leveraging internal and external data.
2. Find relationship between different data sources, for instance, the salary, the credit scoring and the propensity to buy.
3. Produce dynamic rules/ algorithm.
4. Not only make more business value, but also be a tool for risk assessment.

The difficulties that we faced during the development of the Predictive Underwriting are firstly the technical issues, and secondly at the stage of collecting the data supporting Predictive Underwriting, we aren't really sure about the predictive power of an external variable before adding it into the model, and of course, external data cost money, even some sources of internal data cost money as well due to the need of re-planning the IT activities in order to gather all data available at the entities. Based on the previous work on the development of predictive underwriting, we can guess the following sources of external data that could potentially contribute to the predictive modeling:

1. Personal Information: marital status, income level, worth, savings, investment, home value, mortgage value, address, education level, number of children, ...
2. Medical history: can be achieved from the *Medical Information Bank*, other sources of data are available as well such as the information from the vital card, ...
3. Life style/ behavior: hobbies, diet, sport practicing, vehicle type, ...
4. Insurance history: distributor, total premium placed with company, total claim paid by company, seniority, ...
5. Banking information: loan history, credit scoring, seniority, credit card information, ...

2.5 Comparison of Traditional Underwriting and Predictive Underwriting

At the meantime, the advantages of the traditional underwriting are its "relative" accuracy and its ease to explain to the customers. But the processing the underwriting in the traditional way has many inconveniences:

1. It takes time to arrive to the final decision in many cases.
2. It costs the company to the underwriters.
3. It costs the company in the cases that the clients have to take some additional medical tests.

That's the reason why the underwriting's evolution is moving toward the Predictive Underwriting. The advantage of the Predictive Underwriting is that it doesn't take time, nor any additional cost to the company to do all the medical tests, or to recruit many underwriters. But the precision of the Predictive Underwriting is still under the question mark. In this thesis, I will provide you a clear view on how to apply some new techniques on the predictive underwriting and how well can

the accuracy reach.

The other advantage of the predictive underwriting comparing to the traditional underwriting is that it could be more accurate. In this thesis, we will consider the underwriter's decision as the perfect one. But in the future, when the actual claim experience is available, we can use predictive modeling to directly predict the claim's amount. Once we manage to arrive to the desirable accuracy rate, the predictive underwriting can even be better than the traditional underwriting because it reflects directly the claim's amount. But this point will not be in the scope of this thesis because of no data available.

Chapter 3

Data presentation and descriptive statistics

3.1 Type of data used in Predictive Modeling

Recall that AURA is an automated underwriting solution powered by RGA's underwriting expertise. AXA decided to buy AURA from RGA and first applied it to automate the underwriting process in Italy (AXA MPS). So the first available data are from AURA. For the confidential reasons, the data used were modified by adding some coefficients in the manner that doesn't affect the results of the study.

In general, there are 3 kinds of data supporting the predictive underwriting:

1. **Internal data**: data comes from AURA, data contractually captured by the distributors. In the scope of this study, we use the data gathered by AURA, and other data available at the data warehouse of AXA MPS (Italy), for instance, the information about the price, claim history, etc.

2. **External data**: Freely available data, data from AXA's partners. In Italy, the easiest data that we can find without extra fee is from the Italian National Institute of Statistics (ISTAT)'s website. They produce the high quality statistics about the regional data on health, for instance: health status, life expectancy, mortality, life styles, health expenditure, health care demand, household and education, ... (all by region). These are the common and very good predictors to add into our predictive model.

3. **Third party data**: data that we have to pay if we want to get/to use. For example, credit card information, medical history information, ... This kind of data haven't been used yet in our predictive modeling at the first phase, but in the future it will surely become one of the important sources of data if we can find a partner providing good predictors. In the data lake of AXA today, we have some (almost free) available third party data, such as credit scoring, but for only the cases converted to contract. If we want to use them as a predictor in our predictive model, we have to pay in order to have this for all.

3.2 Some descriptive statistics on AURA's data

We have, at the moment of writing this thesis, about 31 262 applications submitted to AURA. More detailed information about AURA: AURA asked systematically 16 questions, in which 6 are non-medical questions, and the other 10 are the base questions, i.e every client has to give the exact responses to these 16 questions. Based on the responses of these 16 questions, other questions can be triggered. The maximum number of questions that a client has to response recorded at the writing moment of this thesis is 210. Our goal at the first step will be to reduce the number of medical questions. To simplify the writing, we associate each AURA decision with a number varying from 1 to 4:

- 1 - Accepted cases.
- 2 - Accepted cases with extra premium.
- 3 - Referred cases.
- 4 - Declined cases.

Note that all the numbers presented in this thesis are not real because of the confidential reason, it is modified by a distortion coefficient in such a manner that it doesn't affect the results of the study.

Here is the distribution of AURA's decision:

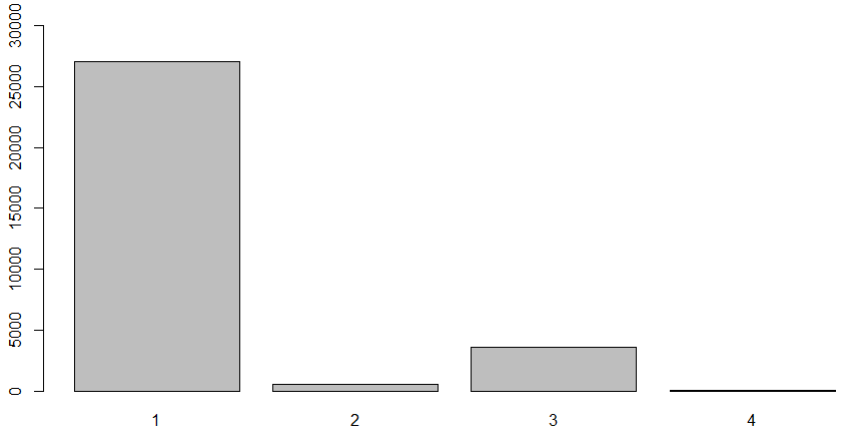


Figure 3.1: Decision of AURA

About 86,7% of the applications are accepted directly by AURA. 11,2% are referred to further analysis by underwriters. Very few cases are declined directly by AURA, because we risk losing good clients, the prospects in this category have clearly very bad health situation making them uninsurable such as cancer at stage 4. There are also not so many cases where extra loading is applied.

Concerning the number of questions, in total of 31 262 clients, 20% of them have to answer the triggered questions, i.e the questions out of the base questions. Among the prospect answering to only the base questions, 94% of them are accepted, 1% is accepted with extra premium, 4% are referred to further examination, very few are refused. Among the prospects answering extra

questions, only 63% are accepted, 1% is accepted with extra loading, 35% are referred to the underwriters, and few of them is refused directly. The acceptance rate reduces because the we normally ask additional questions to the prospect not having the standard health situation. One remark at this stage is that the number of question can be a very good predictor of the underwriting decision, but we can not use this variable because it is not available before we know the underwriting decision.

Contingency table of BMI - Decision

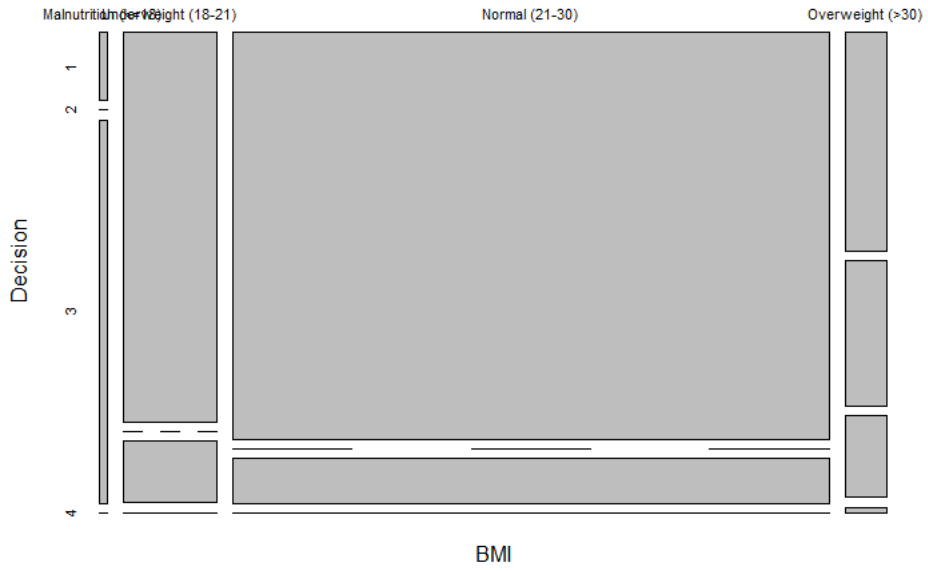


Figure 3.2: Spine plot of BMI and Decision

BMI stands for Body Mass Index and is calculated by:

$$\text{BMI} = \frac{\text{Weight (in kg)}}{\text{Height (in meter)}^2}$$

There are many ways to separate the range of BMI, here we decide to stratify it into 4 ranges in the most common way. The first stratifying has 3 ranges:

- Underweighted range for the prospect having the BMI below 21
- Normal range for those who have the BMI between 21 and 30, this range has in general the good health situation, and should have the highest acceptance rate.
- Overweighed range for the people having the BMI greater than 30

We stratify at the second time the underweighted range into two sub-ranges:

- Malnutrition range for the prospect having the BMI very low, under 17.
- Underweighted range for the prospect having the BMI in between 17 and 21.

Looking at the cross-table plot of BMI and AURA decision, we can clearly see that:

- The Normal range has the highest acceptance rate, which is coherent with what we expected before.
- The Underweighted range ($17 < \text{BMI} < 21$) has lower acceptance rate comparing to the Normal range, but has the higher acceptance rate comparing to the Overweighed range, which is intuitive. Looking at the distribution of AURA's decision on the Overweighed range, we remark that the proportion of "accept with standard" is low, but the proportion of "accept with extra premium" is relatively high, because many of them don't have good health situation but still insurable under some conditions, such as extra premiums, or exclusion of certain types of risk, ...
- In the Malnutrition range, the "accept with standard" proportion is very low because those people usually don't have good health situation, and especially they recover very low from the illness which make the cost of insurance contract very high. But still, many of them are insurable under some special conditions. However, these prospects don't have high proportion in our portfolio.

Contingency table of Gender - Decision

The visualization of two categorical variables could be done by the *cross table* or *contingency table*. It is a group of rectangles, each representing one cell in the two-way contingency table. The area of the rectangle is proportional with the number of observations in the cell. Here, we produce a mosaic plot of *Decision* and *Gender* in Figure 3.3. We also see that the number of male clients is much higher than that of female clients, which is totally normal because the husband is usually the main financial resource of a family, hence needs to be protected.

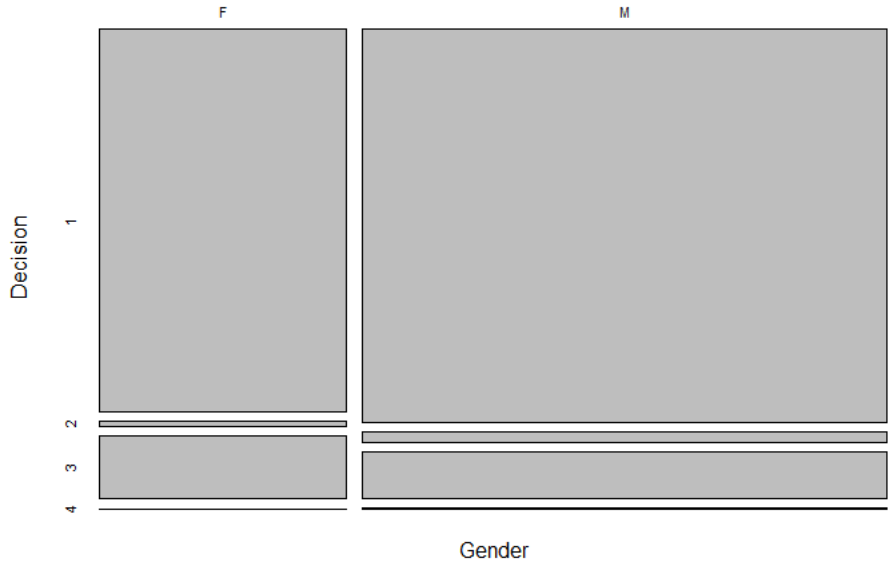


Figure 3.3: Spine plot of Decision and Gender

Another point could be concluded is that the acceptance rate of male clients is statistically higher than that of female clients, which is not intuitive. We want to look at the structure of male - female client to see if our decision is really different for male than female, or it is because of other elements, such that male is generally younger than female, ...

Here is the structure of male - female corresponding to the *BMI*

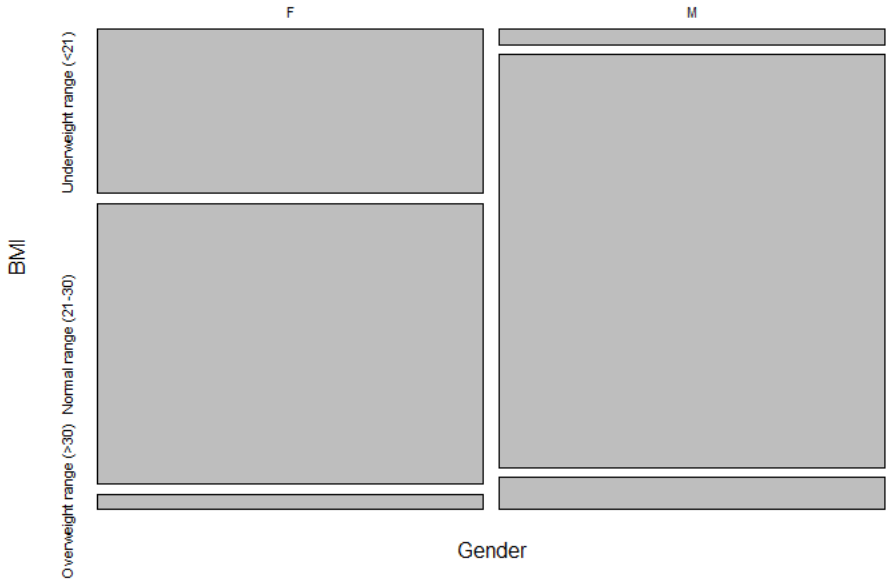


Figure 3.4: Spine plot of BMI and Gender

Here we decided to partition into 3 ranges:

1. Underweight range: those having BMI smaller than 21
2. Normal range: those having BMI in between 21 and 30
3. Overweight range: those having BMI larger than 30

People falling in the second category (normal) have, in general, better health situation than the others. Figure 3.4 shows that the percentage of male being in this category is higher than for female. It might be one of the reasons leading to the high acceptance rate of male.

It is also necessary to look at the internal structure of male - female corresponding to the *smoking status* and the *age*.

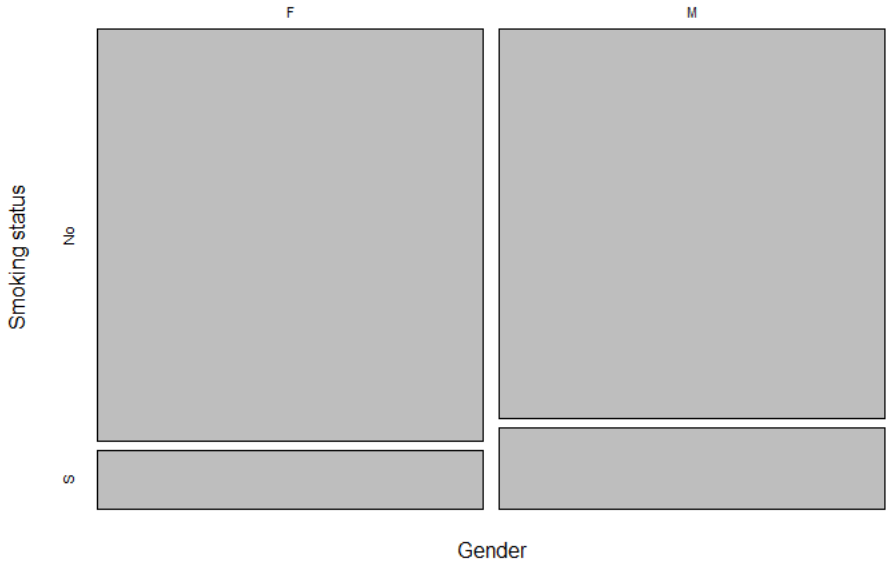


Figure 3.5: Spine plot of Gender and Smoking status

Smokers often refer to the people having the worse health situation than non-smokers. Figure 3.5 shows that the percentage of male smoker is higher than female smoker, which is intuitive. And this should increase the acceptance rate for female, contrary to what we see from the figure 3.3.

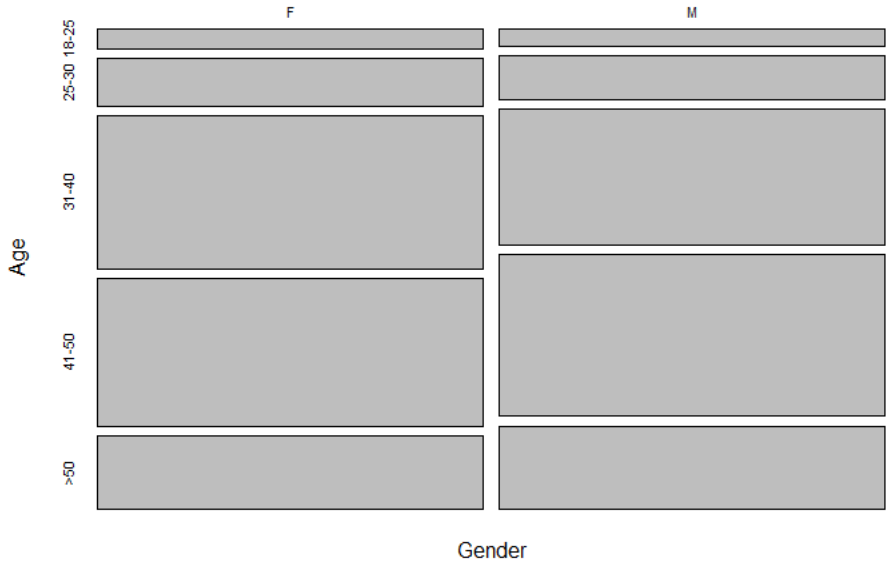


Figure 3.6: Spine plot of Gender and age

Young people often refer to have good health situation. Intuitively they should have high acceptance rate. The figure 3.6 shows that the percentage of young female is higher than young male, illustrated by cells at the bottom of the figure representing the population at 60+ age. This range is the most risky range for the protection product. Based on this, the acceptance rate for female should be higher than male, contrary to the message derived from figure 3.3.

Finally, we can say that the high acceptance rate of male prospects bases mainly on the internal BMI structure of our population. Female prospects have a large amount of people having BMI smaller than 21, which could lead to the problem of malnutrition causing insurance risk in long-term perspective.

Contingency table of Smoking status - Decision

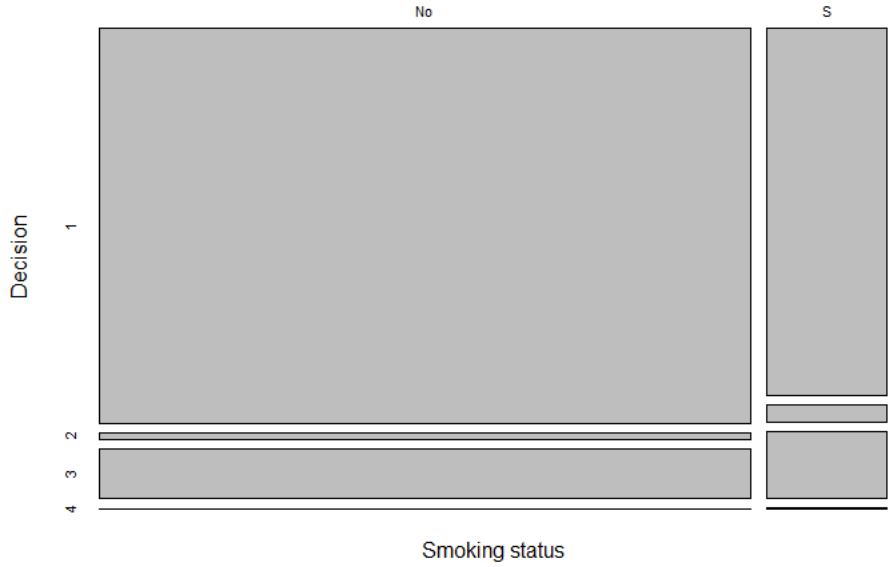


Figure 3.7: Spine plot of Smoke and Decision

It is intuitive that most of prospects don't smoke, and smoking people have lower acceptance rate comparing to non-smoking people.

Contingency table of Age - Decision

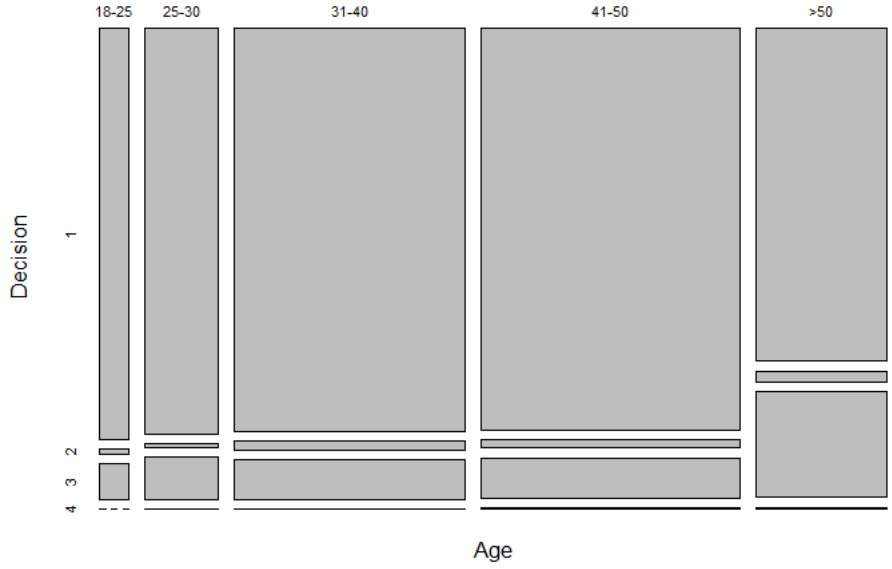


Figure 3.8: Spine plot of Age and Decision

The youngest client is 18 year old. Here we decided to consider 5 ranges of ages, each group has a priori homogenous clients in terms of health expectancy:

- 18 - 25
- 25 - 30
- 30 - 40
- 40 - 50
- 50 +

It is evident that acceptance rate decrease with respect to age. The last group, people being 50+, has very high referred rate, because they usually answer yes in the medical questions concerning the health problem asked by AURA. Hence we need further examinations by the underwriters before making the final decision.

3.3 External data

The easiest external data that could be found is on the website of Italian National Institute of Statistics (Istat.it) [1]. On this website, we can easily find a lot of available data supporting our predictive model. In our protection business, we decided to use the information such as:

- Health and risk factors: health status, life expectancy, mortality, life styles, prevention
- Resources and health care demand: health expenditure, health care supply, health care demand
- Social and economic context: demographic features, fertility and abortion, households and education, environment and territory, employment and poverty.

Those data are available at *Istat* only in form of regional data. Taking health status as an example of external data, we have the information about the number of people reporting chronic conditions: allergic diseases, diabetes, cataract, hypertension, myocardium infarct, ... for instance, regarding hypertension problem, if we take randomly 100 people in Italian population, in Umbria region, we will have in general 16 people having hypertension, this number in Toscana region is only 11,4. In order to include an external data into our model, we need to know at least in which region a prospect comes from. Then we create a new explanatory variable. Always taking example of hypertension problem, the new variable will be named *hypertension* and will take value of 16 for the prospect coming from Umbria region, and will take value of 11,4 if the prospect comes from Toscana region.

This is just the first form of external data, which is not specifically associated to a specific individual. Hence some data will not bring much predictive value to our model because many people have the same value. The more sub-regions we have in the external data, the more predictive value is added to the model.

[1] a public research organization. It has been present in Italy since 1926, and is the main producer of official statistics in the service of citizens and policy-makers

Part II

Machine Learning, Predictive Underwriting and Client Scoring

Predictive Underwriting is not far from now thanks to Big Data and its tremendous potential useful applications for business. Predictive Underwriting is not only a tool to make more business, but also a tool for the risk control. By reducing the number of medical questions, but delivering more or less the same decisions, it can be used to reduce the underwriting's circle time, make our process simpler and hence more competitive comparing to the competitors. It can be used also for detecting the changes of rules in our automated tool, our any abnormal change in the structure of our portfolio.

Machine learning is used to give each prospect an associated *score* corresponding to the information collected from each of them. Ideally, the prospect with better health expectation has higher score. In this thesis, we chose to develop two learning algorithms. The first one is *logistic regression*, which is proved to be a good candidate for the interpretation but is not preferred in term of prediction accuracy. The second one is a black-box algorithm - *Random Forest*. Contrary to *logistic regression*, *random forest* is not a relevant tool for interpretation of models but is proved to be extremely efficient in term of finding hidden patterns giving good predictions.

How can machine learning be used in order to create predictive underwriting models? By using all possible data from AURA's cloud to predict the underwriting decision, we can already see if it is possible to reduce the number of medical questions. By integrating external data into our predictive model, we can improve the performance of our model, and eventually reduce again the number of medical questions.

At the redaction time of this thesis, we had already some applications of predictive modeling. The first one is to:

1. Reduce the number of medical questions in the questionnaire.
2. Partially integrate predictive underwriting into automated underwriting.
3. Give the importance level of each question.

The second application is in the telemarketing approach, where Machine Learning enables us to sort the prospects according to the probability of being accepted, and decide to do telemarketing for only the prospects having the best chance of being accepted. We can eventually much simplify the underwriting process for those prospects.

Chapter 4

Logistic Regression

In this chapter, we will develop one of the most fundamental algorithms: *logistic regression*. The principle of logistic regression is based on the generalized linear model, but logistic regression is used when the response variable is categorical (could be dichotomous or polychotomous). We will first recall the main principles of generalized linear models, then build the binomial logistic regression by seeing how to fit the model, how to test if the model is significant, what is the confidence interval of the estimation, ... After building the binomial logistic regression for the response variable having only two categories, we will extend our work to deal with the response variable having more categories.

After the simple introduction, we will see how can we build the best model, how to choose the relevant explanatory variables. I will introduce you also the different ways, different techniques to evaluate the model, for instance, some statistical tests on the significant of the model, to see the predictive power of the model on new observations, the best predictive model should be the model giving the best performance on the prediction of future outcome, some of the techniques that will be introduced are ROC curve, Cross validation, ...

4.1 Introduction

Logistic regression is a multivariate model commonly used in epidemiology along with multiple linear regression, Poisson regression and Cox model. It is used when the dependent variable (noted Y) is qualitative, usually binary. The explanatory variables (or independent variables, noted X_i) can be moreover qualitative or quantitative. The dependent variable is usually the occurrence or not of an event (illness or other), here we use logistic regression to predict the underwriting decision. The independent variables are those that may influence the occurrence of the event (underwriting decision), that is to say the variables measuring exposure to a risk factor or protective factor, or variable representing a confounding.

The major advantage of this technique is the capacity of quantifying the strength of correlation between each independent variable and the dependent variable, taking into account the effect of other variables included in the model (adjusted measure).

This method is relatively simple to understand and to apply; its results can be easily interpreted as directly related to other methods. The coefficients estimated by the model are indeed mathematically related to the odds (or odds ratio) which represents the strength of the correlation between a risk factor and the underwriting decision, anyway it remains just an approximation of the risk relative. Logistic regression is a good method for searching risk factors or protective factors affecting underwriting decision. However, we should not forget that it is still a mathematical simplification of a complex phenomenon, it theoretically depends on the assumptions, which is frequently unchecked by researchers who apply. Logistic regression is different from the Cox model because it does not allow the inclusion of censored data (that is to say taking into account the individual observation time).

We will first briefly give a theoretical definition of the logistic model and discuss about the essential points of its practice and its interpretation.

4.2 Generalized Linear Model

The traditional linear model describes the relationship between a dependent variable Y, and a set of predictor variables X's such that:

$$Y_i = b_0 + b_1 X_{i,1} + b_2 X_{i,2} + ... + b_p X_{i,p} + \epsilon_i \qquad\qquad i = 1, 2, ..., n$$

In this equation b_0 is the regression coefficient for the intercept and the b_i ($i \geq 1$) are the regression coefficients (for variables 1 through p) computed from the data. Linear models are described as make a set of somewhat restrictive assumptions:

- Dependent variable Y is in normal distribution and conditioned on the value of predictors.
- A constant variance, regardless of the predicted response.

The advantages of linear models with the above restrictions are:

- An easy-to-interpret model form
- Relatively simple computations
- Readily analyzed to determine the quality of the fit

Generalized Linear Modeling (GLM) extends linear regression models to both non-normal distributions and linear transformation (transformation of linearity). It relaxes these restrictions. They adjust responses that violate the linear model assumptions by introducing several concepts:

1. A *random component*: it is represented in the response variable Y_1, Y_2, ..., Y_n with densities belong to the exponential distribution family (gamma, exponential, poisson distribution,...) which can be written as:

$$f(y, \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right)$$

where:

- $\theta \in \mathbb{R}$: canonical parameter.
- $\phi \in \mathbb{R}$: dispersion parameter.

- a: a non-zero real function.
- b: a 2 time-derivable real function.
- c: a $\mathbb{R}^2$-function.

2. A *deterministic component*: for each Y_i $(i = 1, .., n)$, the value of an p-uplet $(X_{i,1}, X_{i,2}, ..., X_{n,i})$ of the variables describing Y_i is known. The vector $X_1 = \big(X_{1,1}, ..., X_{1,n}\big)', ..., X_p = \big(X_{p,1}, ..., X_{p,n}\big)$ are the explanatory vectors.

3. The *link functions* model responses when a dependent variable is assumed to be related to the predictors in a non-linear relationship. These functions, and there are many, transform a target range so the simple form of linear models can be maintained. The link function g is strictly monotone, defined on $\mathbb{R}$ such as:

$$g_n\big(\mathbb{E}(Y)\big) = \underbrace{\beta_0 + \beta_1 X_1 + ... + \beta_p X_p}_{\textit{Score or Linear Prediction}}$$

   Where g is the function from $\mathbb{R}^n$ to $\mathbb{R}^n$ defined by: $g_n(x_1, ..., x_n) = \big(g(x_1), ..., g(x_n)\big)$

4. a *variance functions* used, which express the variance as a function of predicted response. This allows responses with variances that are not constant.

4.3 Dichotomous logistic regression

The regression model that we see usually is the model aiming to predict a continuous outcome based in the function of the continuous predictors. But sometimes, it is possible that we want to predict a binary outcome thanks to one (or several) independent continuous (or categorical) variable(s), this is what logistic regression can do.

Suppose that we want to predict a binary (or dichotomous) dependent variable. By using the regression technique, we could estimate the probability of success for each observation $\mathbb{P}(Y = 1 \mid x_i)$, the probability of failure could be derived directly by $\mathbb{P}(Y = 0 \mid x_i) = 1 - \mathbb{P}(Y = 1 \mid x_i)$.
The problem here is that the linear regression produce the values which are unavoidably out of the range $[0, 1]$ which could not be interpreted as the probability. In fact, linear regressions predict the continuous value on $\mathbb{R}$, and we want only to predict the value within the interval $[0, 1]$

4.3.1 Introduction and fundamental hypothesis of the logistic regression

Suppose that we have a population of n independent observations $(y_i, x_i)_{i=1,...,n}$ (the value of the random variable (X, Y)) where y_i denotes the value of the dichotomous outcome Y, and $x_i = (x_{i,1}, x_{i,2}, ..., x_{i,p})$ is the value of the random vector $X = (X_1, X_2, ..., X_p)$ for the i^{th} observation. y_i takes only two categorical values coded as 0 and 1 (for instance 1 for the accepted cases and 0 for the referred cases). If the variable X_i is interval scale, there is no problem when including it in the model. But if some of them are discrete, or nominal scale such as *smoking status*, *sex*, *occupation*,... we can not simply include them in the model because it makes no sense. In this cases, the method commonly used is to express those variables by a set of *dummy variables*. For example, for the independent variable *sex* having 2 categories *male* and *female*, we need one *dummy variable*, let's say D, taking the value of 0 if the prospect is male, and 1 otherwise. The number of necessary

dummy variables is equal to the number of categories minus 1.

For an individual i corresponding to ω, the value observed is $\big(Y(\omega), X_1(\omega), X_2(\omega), ..., X_p(\omega)\big)$. The probability of being accepted for this individual is $\mathbb{P}\left[Y(\omega) = 1\right] = p(\omega)$. This probability can be easily estimated by $\frac{n_1}{n_0}$ where n_1 is the number of accepted cases and n_0 is the number of referred cases.

We denote $\pi(x)$ the *conditional mean* of Y given x: $\mathbb{E}[Y \mid x]$. Let's remark that $\pi(x)$ is also the *posterior probability* of being accepted, because:

$$\begin{aligned}\mathbb{E}[Y \mid x] &= 1 \cdot \mathbb{P}(Y = 1 \mid x) + 0 \cdot \mathbb{P}(Y = 0 \mid x) \\ &= \mathbb{P}(Y = 1 \mid x)\end{aligned}$$

This is what we are searching to predict in the logistic regression, *posterior probability* of being accepted of the individual ω, i.e the conditional probability that Y equal to 1 given x: $\mathbb{P}\left[Y(\omega) = 1 \mid X(\omega)\right] = \pi(\omega)$.

The model with p independent variables and the j^{th} is categorical (discrete) is written as:

$$\ln\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right] = \beta_0 + \beta_1 x_{i,1} + ... + \sum_{l=1}^{k_j - 1} \beta_{jl} D_{jl} + ... + \beta_p x_{i,p} \quad i = 1, 2, ..., n \tag{4.1}$$

Where $(\beta_0, \beta_1, ...\beta_p)$ are the parameters that we need to estimate. If there is no ambiguity, for simple expressions, we will in general suppress the summation and double subscripting needed to indicate when it is a categorical variable. The model is written under the matrix form as:

$$\ln\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right] = x_i \cdot \beta \tag{4.2}$$

With $x_i = (1, x_{i,1}, x_{i,2}, ..., x_{i,p})$ and $\beta' = (\beta_0, \beta_1, ..., \beta_p)$. The link function chosen here is the *logit* function. The reason why we chose this function is its capacity of transforming any from $(-\infty, +\infty)$ to a number inside the range $[0, 1]$ which can be interpreted as the probability.

Another advantage of logistic regression comparing to linear regression includes the conditional probability of the outcomes. In the linear regression, we assume that the dependent variable $y = x \cdot \beta + \epsilon$ where ϵ is the error supposed to be independently identified distributed following the normal distribution $\mathcal{N}(0, \sigma^2)$, the parameter of variance σ^2 is constant leading to the constant variance of the dependent variable. It is not the case in logistic regressions where we can express the value of the dependent variable Y given x as: $y = \pi(x) + \epsilon$. Recall that y can take only two values 0 or 1, which means that ϵ could take only two values:

$$\epsilon = \begin{cases} 1 - \pi(x) & \text{with probability } \pi(x) \\ -\pi(x) & \text{with probability } 1 - \pi(x) \end{cases}$$

Hence, ϵ follows the Bernoulli distribution with mean zero, and variance $\pi(x)(1 - \pi(x))$ that allows the response variable have the variance not constant.

4.3.2 Fitting the model by the maximum likelihood estimation

The most common method of estimating the parameters of logistic regressions is *maximum likelihood* which searches for the value of β maximizing the probability of observing the data set. In order to apply this method, we have to firstly determine the distribution of $\mathbb{P}(Y \mid X)$. In fact, Y is an binary variable taking its values in $0, 1$:

$$\mathbb{P}(Y \mid x_i) = \pi(x_i)^{y_i} \cdot \big(1 - \pi(x_i)\big)^{1-y_i} \tag{4.3}$$

Thanks to the independent assumption of the variables, it is followed that:

$$\begin{aligned}
l(\beta) &= \prod_{i=1}^{n} \pi(x_i)^{y_i} \left(1 - \pi(x_i)\right)^{1-y_i} \\
\Rightarrow L(\beta) = ln\left[l(\beta)\right] &= \sum_{i=1}^{n} \left\{ y_i \ln(\pi(x_i)) + (1 - y_i) \ln(1 - \pi(x_i)) \right\} \\
&= \sum_{i=1}^{n} \left\{ y_i \ln \frac{e^{x_i\beta}}{1 + e^{x_i\beta}} + (1 - y_i) \ln \frac{1}{1 + e^{x_i\beta}} \right\} \\
&= \sum_{i=1}^{n} \left\{ y_i \cdot x_i\beta - \ln(1 + e^{x_i\beta}) \right\} \\
&= \sum_{i=1}^{n} \left\{ y_i(\beta_0 + \beta_1 x_{i,1} + ... + \beta_p x_{i,p}) - \ln(1 + e^{\beta_0 + \beta_1 x_{i,1} + ... + \beta_p x_{i,p}}) \right\}
\end{aligned} \tag{4.4}$$

Through a log-transformation, a vector $\hat{\beta}$ maximizing the likelihood will also maximize the log-likelihood. The first step to find that value is to differentiate $\mathrm{L}(\beta)$ with respect to $(\beta_i)_{i=(1,..p)}$ to have the following *likelihood equations*:

$$\begin{cases} \frac{\partial L(\beta)}{\partial \beta_0} = \sum\limits_{i=1}^{n} \big(y_i - \pi(x_i)\big) \\ \frac{\partial L(\beta)}{\partial \beta_j} = \sum\limits_{i=1}^{n} x_{i,j} \big(y_i - \pi(x_i)\big) \quad j = 1, ..., p \end{cases} \tag{4.5}$$

For the linear regression, the likelihood functions are linear in β leading to an easy solution to solve those equations. But in logistic regression, we face a system of non-linear equations (4.5) which make the resolution very complicated. Until now, researchers are working on this problem but there has not been any analytic solution. To have the solutions to 4.5, we need to use different numerical methods already coded in some statistical software. In other words, the estimation can vary from one software to another due to the use of different computational methods, details can be found at [11] or [12], page 17 (in French).

The solution of the equations 4.5 is called Maximum Likelihood Estimation (MLE), and has several characteristics:

1. Consistency
2. Asymptotic normality
3. Efficiency

Those properties are very important for the inferential statistics (confident interval, significance tests, ...)

Lets first perform a logistic regression on our data where the goal is to predict the dichotomous underwriting decision, meaning:

$$y_i = \begin{cases} 1 & \text{if } x \geq 0 \\ 2 & \text{if } x < 0 \end{cases}$$

Taking the example on our data, where we want to predict the simple underwriting decision taking only 2 values: accept (regrouping the first and the second cases)/ refer (the other cases) coded as 1,0 respectively based on several basic predictors such as: *kind of product*, *extra premium level*, *occupations*, *gender*, *age*, *smoking status*, *height*, *weight*, *BMI*, and *product coverage amount*. The reason why we solve this problem is to answer to the question: is it possible to give an underwriting decision if we only know some basic information? The result of the fitting procedure is presented in the Appendix A.

The maximum likelihood estimates of $(\hat{\beta}_j)_{j=1,..,p}$ is given in the column *Estimate*. For example, the intercept here takes the value of $-49,37$, and the value of $\hat{\beta}_j$ for the predictor *BMI* is of 0,5734 ... And the estimated conditional probability to be referred for the i^{th} individual is given by the following formula using the estimated value of β in the Appendix A:

$$\hat{\pi}(x_i) = \frac{e^{x_i\hat{\beta}}}{1 + e^{x_i\hat{\beta}}}$$

4.3.3 Estimation of the standard errors

In order to estimate the variances and covariances of the estimated coefficients, we need to, first of all, calculate the second partial derivatives of the log-likelihood function:

$$\begin{cases} \frac{\partial^2 L(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{i,j}^2 \pi_i(1-\pi_i) \\ \frac{\partial^2 L(\beta)}{\partial \beta_j \beta_l} = -\sum_{i=1}^{n} x_{i,j} x_{i,l} \pi_i(1-\pi_i) \quad j,l = 0,1,...,p \end{cases} \tag{4.6}$$

where π_i is $\pi(x_i)$. The *observed information matrix* is defined as the $(p+1) \times (p+1)$ matrix containing the negative of the terms given by the equations 4.6. Let's denote this matrix by $\mathbf{I}(\beta)$. The variance-covariance of the vector $\hat{\beta}$ is calculated by the inverse of $\mathbf{I}(\beta)$:

$$\text{Var}(\beta) = \mathbf{I}^{-1}(\beta) \tag{4.7}$$

Notations:

- $\text{Var}(\beta_j)$ the j^{th} diagonal element of $\mathbf{I}^{-1}(\beta)$, corresponds to the variance of β_j.
- $\text{Cov}(\beta_j, \beta_l)$ the $(j,l)^{\text{th}}$ element of $\mathbf{I}^{-1}(\beta)$, corresponds to the covariance between β_j and β_l.

The estimations of the variance and covariance of the estimated coefficients will be denoted by $\widehat{\text{Var}}(\hat{\beta})$, obtained by taking the value of $\hat{\beta}$ in the matrix $\text{Var}(\beta)$. Intuitively, we denote the estimated of variance and covariance in this matrix by $\widehat{\text{Var}}(\hat{\beta}_j)$ and $\widehat{\text{Cov}}(\hat{\beta}_j, \hat{\beta}_l)$, $j,l = 0,1,...,p$. Hence, the estimated standard errors of the estimated coefficients can be calculated as:

$$\hat{\sigma}(\beta_j) = \left[\widehat{\mathrm{Var}}(\hat{\beta}_j)\right]^{1/2} \tag{4.8}$$

[8] introduced an useful formulation helping to calculate the information matrix:

$$\underbrace{\hat{\mathbf{I}}(\hat{\beta})}_{(p+1)\times(p+1)} = \underbrace{X'}_{(p+1)\times n} \cdot \underbrace{V}_{n\times n} \cdot \underbrace{X}_{n\times(p+1)} \tag{4.9}$$

where:

$$X = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix}$$

and

$$V = \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & \dots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & \dots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \dots & 0 & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix}$$

The result of all those calculations are presented in the column headed *Std. Error* in the Appendix A.

4.3.4 Significance test of the coefficients

The goal of the significance test is to prove the role of one or several explanatory variables, to answer the question should we include these variables in our model or not, does the model including the testing variable tell us more about the outcome than a model not including that variable. This is equivalent to see how much an individual variable contributes to the overall model.

The null hypothesis can be written as following:

$$H_0 : \beta_i = 0$$

We can compare two different models, one including the variable that we want to test, and one does not. If the overall model gets significantly better in some sense, then we say that the tested variable is statistically significant, and vice versa. It is important to note that, at this state, we don't consider the quality of the model in the absolute sense (we don't care if the model is overall accurate or not), but in an relative sense (i.e, the question concerned here is if one model is better than the other). The principle of the comparison would be: taking the first model including the variable in question, evaluate how good the model is by comparing the predicted values to the observed values. Then do the same for the second model not including that variable. And finally, compare the goodness of two models above. There are two approaches for implementing the test:

1. Test based on the likelihood: this approach is powerful and coherent with how we estimate the parameters. But the disadvantage of this approach is on the computational resources, because at each hypothesis to be tested, we need to re-estimate the parameters, so one more time using numerical optimization presented above.

2. Wald's test: based on the asymptotic normality property of MLEs. The principle advantage of this test is that all the necessary information we want is available after fitting the completed model, so the result can be calculated directly, don't take much resource in the hardware. But the disadvantage is that this test is conservative and tends to favor the null hypothesis.

Test based on the likelihood

Just like in the linear regression, we want to evaluate the goodness of the model based on the square of the distance between the observed and the predicted values. In logistic regression, it is not as easy as the linear regression due to the fact that the outcome here is dichotomous. An equivalent way to perform the test is to base on the log-likelihood function defined in (4.4). Let's consider the following ratio:

$$\begin{aligned} D &= -2\ln\left[\frac{\text{likelihood of the fitted model}}{\text{likelihood of the saturated model}}\right] \\ &= -2\sum_{i=1}^{n}\left[y_i\ln\left(\frac{\hat{\pi}(x_i)}{y_i}\right)+(1-y_i)\ln\left(\frac{1-\hat{\pi}(x_i)}{1-y_i}\right)\right] \end{aligned} \tag{4.10}$$

This quantity is called *deviance*, and the ratio inside the brackets is the *likelihood ratio*. The saturated model is the model containing as many parameters as there are observations. The deviance gives us an idea of the comparison between the predicted values (represented by the fitted model) and the observed values (represented by the saturated model) if we conceptually consider an observed value as also being a predicted value from the saturated model. Because the likelihood ratio is always negative, we multiply its log by minus 2 not only to have an positive number but also to have a quantity whose distribution is known.

Following the previous principle, in order to evaluate the significance of an covariate, we need to compare the deviance of each fitted model, the change of the deviance due to the inclusion of that covariate is:

$$\begin{aligned} G &= D(\text{model without the variable}) - D(\text{model with the variable}) \\ &= -2\ln\left(\frac{\text{likelihood without the variable}}{\text{likelihood with the variable}}\right) \end{aligned}$$

Following [8], in case that we have only one independent variable:

$$G = -2\ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}\hat{\pi}_i^{y_i}(1-\hat{\pi}_i)^{1-y_i}}\right] \tag{4.11}$$

$$= 2\left\{\sum_{i=1}^{n}\left[y_i\ln(\hat{\pi}_i)+(1-y_i)\ln(1-\hat{\pi}_i)\right]-\left[n_1\ln(n_1)+n_0\ln(n_0)-n\ln(n)\right]\right\} \tag{4.12}$$

where:

- $n_1 = \sum y_i$ is the number of referred cases.
- $n_0 = \sum(1 - y_i)$ is the number of accepted cases.

Under the hypothesis $H_0 : \beta_1 = 0$, G follows a chi-square distribution with 1 degree of freedom. [8]. Taking an example where we want to test the significance of the variable BMI of the model above:

- The deviance of the model including BMI is: 699,39
- The deviance of the model not including BMI is: 704,07

Thus, the statistic G is equal to: $704,07 - 699,39 = 4,68$. Note that $G \sim \chi^2(1)$: a chi-square distribution with 1 degree of freedom. The p-value associated with this test is:

$$\mathbb{P}\left[\chi^2(1) > 4,68\right] \approx 0,03 < 0,05$$

The result of the *likelihood ratio test* tells us that BMI contribute significantly to the prediction of underwriting decision using previous model. Before concluding that BMI is a ultimately significant variable, we need to take into account the goodness of the model, the significance of the addition of new predictors, ... These points will be discussed in the subsequent sections

The result of this test is not given immediately in R with the function `glm`, we had to firstly fit 2 models (with and without the variable in question) in order to have the value of the deviance of each model. Then we calculated by hand the p-value. Suppose that we want to calculate the significance level of 100 variables, the work is to fit 100 different models (with and without the variable in question), then calculate the difference between two values of deviance, then calculate the p-value. This process is time-consuming. Fortunately, we have an alternative: Wald's test.

Wald's test

This test is based on the asymptotic normality of the coefficients, i.e , when n is sufficiently large, the vector $\hat{\beta}$ follows the multivariate normal distribution. The statistic of the Wald's test is given by [12] as below:

$$Z_j = \frac{\hat{\beta}_j}{\hat{\sigma}_j} \sim \mathcal{N}(0,1) \tag{4.13}$$

where:

- $(\hat{\beta}_j)_{j=(0,1,\ldots,p)}$ are the MLEs.
- $(\hat{\sigma}_j)_{j=0,1,\ldots,p}$ are the associated estimates of the standard errors of the estimated parameters. How to obtain these values will be discussed formally in the following section. What need to know here is that the values of $(\hat{\sigma}_j)_{j=0,1,\ldots,p}$ are given immediately in R with the function `glm` (in the column headed *Std. Error* in the Appendix A).

The value of the Wald's test statistic is also given by the column head *z value* of the function `glm` in R, which is the quotient of the two previous columns in the Appendix A. Z_j can take the negative values, so the test is bilateral, the p-value of Wald's test can be calculated by calculating:

$$\mathbb{P}\left(|z| > Z_j\right) \tag{4.14}$$

where z denotes a random variable following the standard normal distribution. These values are given at the last column of the result of the function `glm` in R. The detail results can be found at the Appendix A. Based on the result of Wald's test, the following variables are significant: *product type*, *gender*, *age*, *smoking status*, *height*, *weight*, *BMI*. It's important to know that the disadvantage of this test is that it works in an aberrant way, sometimes tends to reject the null hypothesis when the coefficient is significant. [8]

4.3.5 Confidence interval of the estimation

Based on the fact that $(\hat{\beta}_j)_{j=(0,\dots,p)}$ follow asymptotically the normal distribution, we can construct their $(1-\alpha)$ confidence interval for each individual coefficient as:

$$(\hat{\beta}_j \pm z_{1-\alpha/2} \cdot \hat{\sigma}_j)_{j=(0,1,\dots,p)} \tag{4.15}$$

where $z_{1-\alpha/2}$ is the upper $(1-\alpha/2)$ point from the standard normal distribution. The calculation of the $\hat{\sigma}_j$ will be discussed in the following section. For the moment, we use the value provided by the output of R.

4.4 Polychotomous logistic regression

4.4.1 Fitting the model

When the outcome has more than 2 categories ($K > 2$, for instance, the situation where we want to predict the underwriting decisions with more than 2 levels: accept, accept under certain conditions, refer, decline), we talk about the *polychotomous logistic regression* (or *multinomial logistic regression*).

We define the probability of being in the category k of the i^{th} individual as:

$$\pi_(x_i) = \mathbb{P}(Y_i = k \mid x_i) \tag{4.16}$$

subject to $\sum_{k=1}^{K} \pi_k(x_i) = 1$.

Consequently, the likelihood function is expressed as:

$$L = \prod_{i=1}^{n} \left[\pi_1(x_i)\right]^{y_1(x_i)} \times \dots \times \left[\pi_K(x_i)\right]^{y_K(x_i)} \tag{4.17}$$

where:

$$y_k(x_i) = \begin{cases} 1 & \text{if } Y_i = k \\ 0 & \text{otherwise} \end{cases} \tag{4.18}$$

It is the generalization of the binary case when we have the binomial distribution for a dichotomous response variable Y. In order to fit the model, the first thing to do is to fix a baseline outcome, and then express the logit with respect to the baseline outcome (or reference category). We can chose any category as a baseline outcome, there is no impact on the calculation/quality of prediction of the model, the only difference lies on the interpretation of the coefficients. Without loss of generality,

we address the K^{th} category as the reference category. The logit for the k^{th} category can be written as:

$$C_k = \ln \frac{\pi_k(x_i)}{\pi_K(x_i)} = x_i \cdot \beta_k \quad \text{for} \quad k = 1, ..., K-1 \tag{4.19}$$

$$\rightarrow \pi_k(x_i) = e^{x_i \cdot \beta_k} \cdot \pi_K(x_i) \quad \text{for} \quad k = 1, ..., K-1 \tag{4.20}$$

where $\beta_k' = (\beta_{0,k}, \beta_{1,k}, ..., \beta_{p,k})$.

Using the fact that all probabilities must sum to one, we find:

$$\pi_K(x_i) = \frac{1}{1 + \sum_{k=1}^{K-1} e^{x_i \cdot \beta_k}} \tag{4.21}$$

We can use this to find the other probabilities:

$$\pi_1(x_i) = \mathbb{P}(Y_i = 1 \mid x_i) = \frac{e^{x_i \cdot \beta_1}}{1 + \sum_{k=1}^{K-1} e^{x_i \cdot \beta_k}}$$

$$\pi_2(x_i) = \mathbb{P}(Y_i = 2 \mid x_i) = \frac{e^{x_i \cdot \beta_2}}{1 + \sum_{k=1}^{K-1} e^{x_i \cdot \beta_k}}$$

$$\cdots$$

$$\pi_{K-1}(x_i) = \mathbb{P}(Y_i = K-1 \mid x_i) = \frac{e^{x_i \cdot \beta_{K-1}}}{1 + \sum_{k=1}^{K-1} e^{x_i \cdot \beta_k}}$$

In order to estimate $(K-1) \times ((P+1)$ coefficients, we need to optimize the log-likelihood:

$$LL = \sum_{i=1}^{n} y_1(x_i) \ln \left[\pi_1(x_i)\right] \times \cdots \times y_K(x_i) \ln \left[\pi_K(x_i)\right] \tag{4.22}$$

Here we are in the same situation as in the binary cases, there is no explicit solution of the equations. We can calculate the approximate value of β_k by the numerical algorithms such as the algorithm of Newton-Raphson. This is done via the function `multinom` of the package `nnet` of R.

Here are the short-cut results of the `multinom` function on the data from AXA Italy (more details can be found at Appendix **??**):

	Intercept	Gender - Male	Age	Smoke	Height	$\cdots$
2 (accept with extra premium)	-24,252	-0,043	-0,017	0,692	-0,114	$\cdots$
3 (refer)	-19,816	0,145	0,113	0,489	0,174	$\cdots$
4 (decline)	32,736	-5,036	-0,069	-0,235	-0,538	$\cdots$

Here we chose the first category (accepted cases) as the reference category, and we calculate 3 vectors of (β_k) corresponding to the three categories left. One can chose any different baseline outcome, nothing will change in term of calculation or prediction.

4.4.2 Significance test

The significance test of the coefficients in the multinomial logistic regression is more complicated than in the binary logistic regression, because we can have eventually multiple possibilities. And especially, the consequence is not the same. For instance, if a variable is significant in all of the $K-1$ logits ($K>2$), then there is no problem to conclude that this variable is significant. But when the statistic test shows that a variable is significant in at least one logit, and insignificant in the others, the conclusion is not evident at all.

Other type of test could be interesting as well, such as the test on the equality of the coefficient for several (or all) logit equations. This test is not for the significance of the coefficients but to test if the variable questioned has the identical impact on the different categories of the response variable.

Like in the binary case, we have two methods to perform the test. One is based on the log-likelihood function, and the other, Wald's test, is based on the asymptotic normality of the estimated coefficients.

Test based on the log-likelihood function

The null hypothesis is:

$$H_0 : \begin{cases} \beta_{j,k} = 0 & \text{when we want to test the significance in one logit} \\ \beta_{j,k} = 0, \forall k & \text{when we want to test the significance in all logits} \end{cases} \tag{4.23}$$

The first one answers the question: is the coefficient in a specific logit is significant? We can not say anything about its significance in other logits. While the second one goes further and let us know if the coefficients of one explanatory variable are simultaneously null in all logits.

The principle is more or less the same as before in the binary case, the statistic is calculated by taking the difference between the deviances of the regressions, it follows the χ^2 distribution under the hypothesis H_0. The degree of freedom is obtained by the difference between the number of estimated parameters.

$$LR = D_{H_0} - D_M \tag{4.24}$$

where D_{H_0} is the deviance of the model under H_0, and H_M is the deviance of the full model (model including the variable that we want to test). It is proved that $LR \sim \chi^2$ with 1 degree of freedom in the first case, and $K-1$ degree of freedom in the second case.

Wald's test

The idea stays unchanged, the first thing to do, which is a little bit complicated, is to calculate the variance - covariance matrix of the coefficients. The function `multinom` of R enables us to know this matrix under the form of the Hessian matrix. Wald's statistic is thus calculated by:

$$\begin{cases} W_{j,k} = \frac{\hat{\beta^2_{j,k}}}{\hat{\sigma}_{\hat{\beta}_{j,k}}} & \text{for the significance test in one logit} \\ W_j = \beta_j' \hat{\Sigma}_j^{-1} \hat{\beta}_j & \text{for the significance test in all logits} \end{cases} \tag{4.25}$$

where $\hat{\beta}'_j = (\hat{\beta}_1, \cdots, \hat{\beta}_{K-1})$ is the vector of coefficients to be tested, and $\hat{\Sigma}_j^{-1}$ is its variance - covariance matrix. Under H_0, it follows the χ^2 distribution with the degree of freedom equal to 1 in the first case, and equal to $(K-1)$ in the second case.

4.5 Model building strategies

4.5.1 Introduction

In the context of Big Data, we may have an enormous number of independent variables to include in our model, making the possibility of having many possible models, not all are good, not only in term of fitting, but also in term of generalization and prediction. Because some variables could be significant having no linear association with the response variable. The best model is the model having less independent variables possible, but still able to deliver good predictions. Fewer variables means easier interpretation: by excluding the unnecessary variables, we can more easily identify the role of the variables being chosen. Moreover, it saves costs for the business as well because we don't need to spend money collecting the unnecessary information. Another point that we should take into consideration is that a model with fewer independent variables is appeared to be more robust in the generalization (and prediction). In fact, when we have too many independent variables, we will face the famous problem named *over-fitting*, i.e the model fits "too well" the data, hence, in stead of learning from the essential information represented in the population, the model learns "perfectly" the noise as well. Including unnecessary independent variables increases the variances of the estimated coefficients, making them unstable relatively to the data observed, and the model becomes more dependent of the data. This is not good because in the future data, the performance of the model will not be stable. In this situation, we need to define the strategies or methods for choosing the appropriated variables to include in our model. It could also help us to save the cost of buying/collecting unnecessary external data.

The goal is to select as less variables as possible, resulting the "best" model. Thus there are two questions that we will answer in this section:

1. How to select (include/exclude) independent variables?
2. What is a good model?

4.5.2 Variable selection

In the building phase of the model, we should not consider only the statistical/scientific point of view, the independent variables should be chosen also based on experience and based on the meaning of the variable in the reality. Consequently, based on the same strategy of selecting variables, two different people with different backgrounds may end up choosing different models. It is rarely evident to say which one is better.

This choice is essential. It must be based on prior knowledge of the problem to be treated, especially on possible confounding. The logistic model used for predictive underwriting must be based on assumptions and knowledge "causal network" that is woven around an underwriting decision. We must in advance perform a precise multivariate analysis and complete descriptive analysis (distribution of variables, recoding, and / or regrouping ...) then a univariate logistic regression analysis.

The methodologies or criteria for choosing the best model can be different from one statistician to the other, because of the different backgrounds, different interpretation of the clinical importance of each variables, but the final goal is always to find the most parsimonious model that still predict well the outcome. Some people propose to include all possible intuitively relevant variables into the model not considering their statistical significance, in order to better control the confounding. By doing that, the model will fit really (even too) well the data, and will capture not only the general trends, general structures of the data, but will also capture (well) significantly some noise presented in the data. Strictly speaking, the estimated coefficients and/or estimated standard errors are very sensitive to the data, making it unstable over the time especially when we don't have sufficient observations in our database, which is the case here because AXA has just implemented AURA in Italy for more than 5 months. This will be a big problem in the prediction phase because the noise in the past data is not necessarily similar to the noise in the future data, which lead to the good model in term of fitting the actual data, but bad model in term of predicting the outcome. Once again, the problem of *overfitting* arises here, the symptom of overfitting is usually associated with the unrealistically large estimated coefficients and/or estimated standard errors. That's the reason why we didn't decide to drive our work in this way, but we processed by the following steps.

Choosing potential variables

The goal of this step is to choose the relevant variables to be taken into account, from a set of available predictors. Intuitively, we will choose the ones that have strong associations with the response variable, and exclude the ones that have no (or minor) relationship with the response variable. We begin our work with the univariable analysis of each variable. The way to treat the different kinds of variables is not the same and is described as follow:

1. For the nominal/discrete independent variables: we want to test the association of the outcome vs the independent variable, it can be done by a contingency table of the response variables vs K levels of independent variables. The likelihood ratio chi-square test with $K-1$ degrees of freedom is exactly equal to the value of likelihood ratio test for the significance of the coefficients for the $K-1$ design variables in a univariable logistic regression model containing only that single independent variable in logistic regression. [8].

2. For the continuous independent variables: in order to measure the association with the response variable, we can fit a univariable logistic regression containing only the questionable variable in order to know: the estimated coefficients, the estimated standard errors, the likelihood ratio test for the significance of the coefficients, the univariable Wald's statistic.

Here is the table of independence test of the categorical/discrete independent variables:

Predictor	Test	X-square	DF	p-value
Gender	Chi-square	80,8	3	$< 2,2e-16$
Smoke	Chi-square	103,09	3	$< 2,2e-16$
Occupation	Chi-square	4136,8	345	$< 2,2e-16$*
Sport avocation	Chi-square	428,92	60	$< 2,2e-16$*
Has an	Chi-square	375,08	3	$< 2,2e-16$*
Are you presently	Chi-square	929,9	3	$< 2,2e-16$*
Do you currently	Chi-square	127,83	3	$< 2,2e-16$*
Has your	Chi-square	6282,1	3	$< 2,2e-16$*
Do you intend	Chi-square	945,38	3	$< 2,2e-16$*
X5 year 1	Chi-square	2613,4	3	$< 2,2e-16$*
X5 year 2	Chi-square	154,4	6	$< 2,2e-16$*
Do you receive	Chi-square	269,98	6	$< 2,2e-16$*
Apart from	Chi-square	2175,6	3	$< 2,2e-16$*
Product	Chi-square	513,65	18	$< 2,2e-16$*

Table 4.1: Chi-square test of Independence for categorical independent variables

As the p-value are all $< 2,2e-16$, which is significantly smaller than the 0,05 significance level, we reject the null hypothesis that the dependent variable is independent of the independent variables. This is intuitive because the underwriting decisions of AURA are delivered based on the answers of the questionnaire.

Those cases having the sign * at the right hand side of the p-value are the cases having the following warning message:

```
Warning message:
In chisq.test(table(survey$Smoke, survey$Exer)) :
  Chi-squared approximation may be incorrect
```

It gave the warning because of the small cell values in the contingency table leading to very small expected values of p, so the value appeared can be not exact. To avoid this problem, we can regroup the different categories in a reasonable way to avoid the small cell values, then apply the `chisq.test` function against the new regrouped data instead. But it is not of our interest because at this step, we only want to test if the independent variable has a relationship with the dependent variable or not. We are satisfied with the fact that p-value is smaller than the significance level 0,05.

And here is some information concerning the fitting of the univariable logistic regression with respect to each continuous independent variable:

Variable	Coefficients			Standard errors			p-value			Deviance
Categories	2	3	4	2	3	4	2	3	4	
Age	0,03	0,059	0,03	4,5e-03	2,2e-03	0,02	7,5e-08	0,0e+00	6,0e-02	9846,4
Height	-2,5e-0,4	-0,02	-0,04	2,9e-04	2,5e-03	9,2e-04	3,9e-01	1,2e-13	0,0e+00	10203
Weight	0,15	-8,1e-04	0,15	4,4e-03	1,7e-03	8,3e-03	0,00	0,64	0,00	9226
BMI	0,69	0,03	0,71	0,02	6,9e-03	0,03	0,0e+0	1,1e-06	0,0e+0	8818,4
Sum CI	7,3e-08	-2,6e-06	5,5e-06	1,9e-06	9,9e-07	4,2e-06	0,97	7,4e-03	0,19	2171,6
Sum Life	1,6e-06	1,7e-06	4,3e-07	4,4e-07	2,4e-07	2,1e-06	2,7e-04	1,2e-12	8,3e-01	9986,8
Sum LTC	3,4e-04	-1,3e-05	1,6e-04	5,1e-04	1,9e-04	7,6e-04	0,51	0,95	0,84	679,2

Table 4.2: Univariable analysis for continuous independent variables

We can easily see that neither the estimated coefficients nor the estimated standard errors are unrealistically large. Additionally, the p-values of the Wald's test in most of the cases are smaller than 0,05. Except for the variable Sum LTC (sum assured of the LTC product), the reason is that there are only few contracts in our database are LTC contracts. But as mentioned before, variable selection is not only based on the statistical point of view, it's based also on the clinical meaning of each variable. Seeing that the coverage amounts for CI and Life are both significant, so we decided to keep the variable *Sum LTC* for the next steps. This is not dangerous because we can always exclude this variable in the following step, but by excluding this variable at this step, we take the risk of excluding a significant variable from the model.
In general, we include in this initial model the independent variables whose degree of significance is less than 0,2 in the univariate model because they can then be confounders or be influenced by other variables in a multivariate model and become hence significant. We can also eventually include experienced relevant variables, the obvious confounders whose univariate analysis would however not lead to an p-value less than 0,2. Sometimes it can be helpful to create several fictitious binary variables (dummy variables) to represent each item of polytomous variables. Finally, in some cases, the correlation between two variables that should be entered in the initial model can be important (they both provide the same information), then the coefficient of the models can not be estimated correctly. We speak in this case about the collinearity between independent variables. To avoid those problems, it is particularly important to select the initial variables in the experience and statistic point of view.

4.5.3 Selection of the preliminary model

Once the first step of preliminarily selecting variables is done. We go to the second step, in which all the variables whose univariable test gives a p-value smaller than $\nu = 0,25$ will be selected as a candidate, along with all the variables of known clinical importance. The choice of $\nu = 0,25$ is arbitrage, if ν is too small, we are too selective at this step, thus sometimes there is the risk of failing to include potentially significant variables. Because, a variable which is not significant in univariate model can eventually become significant when going together with the other variables in the multivariate model. So, we need to increase the threshold 0,05 in order to better capture all the possible significant independent variables. But at the same time, we shouldn't increase too much the value of ν because if we do so, when ν is big, and when we have in the model thousands of independent variables, there could be too many independent variables to test in the second step, which will make the work less effective and time-consuming.

Several approaches:

Backward selection

1. Begin with the model containing all the variables selected from the first step.
2. Estimate the parameters of the logistic regression.
3. Choose among the coefficients the one having the lowest p-value of the Wald's test.
4. Verify if it's significant or not by comparing with the confidence level $1 - \alpha$ that we decide personally. If p-value $\leq 1 - \alpha$, we keep the variable, and end the process here. If not, p-value $\geq 1 - \alpha$, the variable will be excluded from the collection of independent variables. Then we continue the process by going back to 2.,

Forward selection

Contrary to the backward selection, one can progressively include each independent variable (forward regression) by taking out those who are not significant or become not insignificant after including a variable. The details of forward selection are presented as follow:

1. Fit the initial model with only a constant, without any independent variable.
2. Among the independent variables identified as associated with the dependent variable in the first step, detect the one maximizing the statistical test when adding it to the existing model.
3. Verify if that variable is significant or not (if p-value $\leq 1 - \alpha$. If yes, include that it in the existing model, then estimate the parameters of the new model. If there are still the variables not checked, going back to the 2., if the variable is not significant, it should not be selected. The process continues until there is no variable left.

Stepwise selection

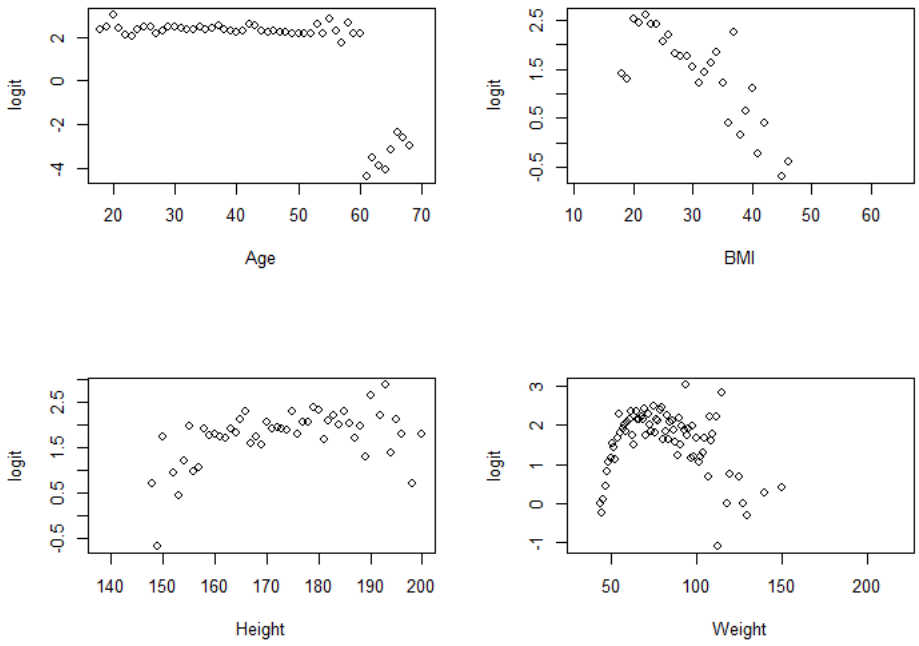
Stepwise method involves selecting variables either for inclusion or exclusion. It is the mixed version of backward and forward selection, and there exists two versions: 1. forward selection with a test for backward elimination, and 2.backward elimination with a test for forward selection. The stepwise method is used as the reference method in many implementation packages for the automated variable selection because of its speed. But recently, it is proved that the stepwise is not effective and fails to select the good model especially in the presence of collinearity.

Compared to the forward selection, the backward elimination, which consists of including all selected variables and progressively remove non-significant one, adds an interesting property: it takes into account the combinations of variables. Because, sometimes a variables is significant only in the presence of some other variables. As the backward strategy begins with the model containing all the selected variables, it will not leave out this kind of situation. Furthermore, the calculation time for the backward elimination is smaller than for the forward selection.

4.5.4 Selection of the final model

Once the preliminary model is reached, i.e the model containing all essential variables, a look in details at each independent variables is necessary. The first thing to do is to check the linearity in the logit assumption of the model. The figure below illustrates the relationship between some

continuous variables and the logit. We can clearly see the linear relationship between *BMI* and the logit. But for the *Age*, there is a big difference in the logit between two groups: people being at most 60 years old, and people being more than 60 years old.



The conclusion of this stage is that the variable *Age* violates the linearity assumption. However, we can easily fix it by transforming this variable into the categorical variable with respect to 2 groups identified previously.

$$\text{Age (categorical)} = \begin{cases} 0 & \text{if Age} < 60 \\ 1 & \text{if Age} \geq 60 \end{cases} \tag{4.26}$$

One of the last things to do in the building phase of the model is to check for the multicollinearity between independent variables. This is the phenomenon in which at least two independent variables are highly correlated. One of the solution for detection of multicollinearity is to look at the changes in the estimated coefficients when a explanatory variable is added or deleted in the backward/forward/stepwise logistic regression. When we have lots of predictors, it can be problematic. Another way to detect the multicollinearity is to calculate the *Variance Inflation Factor*(VIF). The `car` package in R implements a slightly different VIF calculation called GVIF (generalized VIF), the interpretation stays unchanged.

$$\begin{cases} \text{GVIF} = 1 & \text{i.e No correlated} \\ 1 \leq \text{GVIF} < 5 & \text{i.e Moderately correlated} \\ 5 \leq \text{GVIF} & \text{i.e Highly correlated} \end{cases} \tag{4.27}$$

We attempt to eliminate variables with a GVIF higher than 5.

Predictors	GVIF	Df	$GVIF^{(1/(2*Df))}$
Gender	2,385053	1	1,544362
Smoke	1,035293	1	1,017494
Height	55,941139	1	7,479381
Weight	220,997010	1	14,865968
BMI	121,524256	1	11,023804
Sport.avocation	1,064226	18	1,001731
Has.an	1,000000	1	1,000000
Are.you.presently	1,003525	1	1,001761
Do.you.currently	1,017873	1	1,008897
Are.you.currently	1,308395	22	1,006128
Has.your	1,009044	1	1,004512
Do.you.intend	1,028632	1	1,014215
X5.year.1	1,193664	1	1,092549
Apart.from	1,002639	1	1,001319
Sum.CI	1,111120	1	1,054097
Sum.Life	1,034880	1	1,017291
Sum.LTC	1,082916	1	1,040632
Occupation	1,469269	20	1,009666
Age	1,362432	1	1,167233

Table 4.3: Results of the `vif` function

In this calculation, we excluded some variables known to be not correlated but will affect the calculation of GVIF. Those variables are: *X5.year.2*, *Do.you.receive* and *PRODUCT*, because they correspond to the 2 questions asked only for LTC product, so there are missing values for other products, and Yes/No value for LTC product. The model will understand that these 2 variables are highly correlated (they are not) and the calculation of GVIF values for other variables will be affected.

In our case, three variables lead to the GVIF higher than 5: *Height*, *Weight*, and *BMI*. From the clinical point of view, we eliminated 2 variables *Height* and *Weight* as the information from these variables is somewhat already reflected in the variable *BMI*. The relationship between *BMI* and *Weight* is confirmed by the following graph:

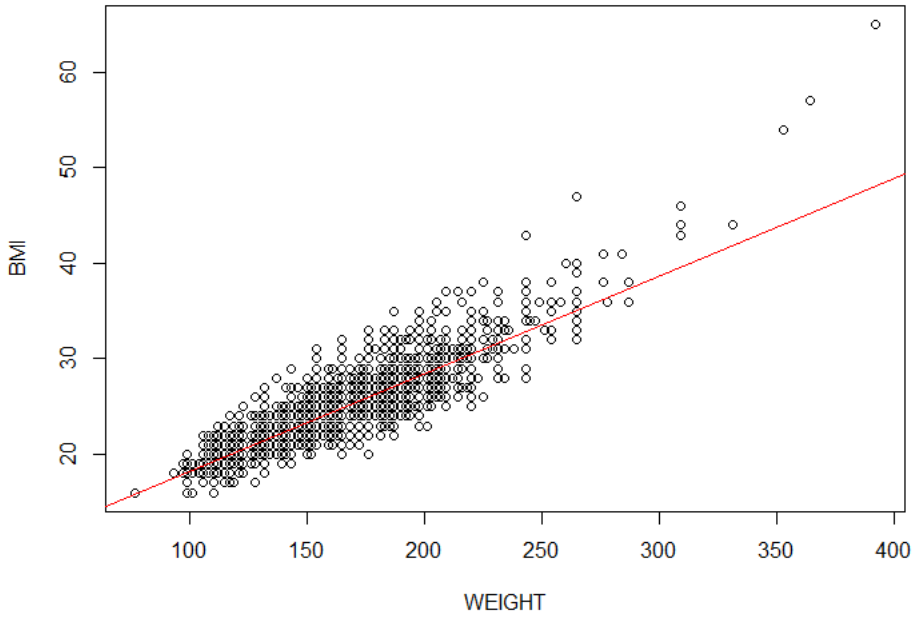


Figure 4.1: Relationship between BMI and Weight

The values of GVIF for each variables after excluding *Height* and *Weight* are presented in the following table:

Predictors	GVIF	Df	$GVIF^{(1/(2*Df))}$
Gender	1,498646	1	1,224192
Smoke	1,032814	1	1,016275
BMI	2,032284	1	1,425582
Sport.avocation	1,064296	18	1,001732
Has.an	1,000000	1	1,000000
Are.you.presently	1,003527	1	1,001762
Do.you.currently	1,013485	1	1,006720
Are.you.currently	1,290501	22	1,005813
Has.your	1,009366	1	1,004672
Do.you.intend	1,023420	1	1,011642
X5.year.1	1,191953	1	1,091766
Apart.from	1,002525	1	1,001262
Sum.CI	1,111317	1	1,054190
Sum.Life	1,035690	1	1,017688
Sum.LTC	1,082900	1	1,040625
Occupation	1,436087	20	1,009089
Age	1,354251	1	1,163723

Table 4.4: Results of the `vif` function after excluding correlated variables

Now all the explanatory variables are not correlated. We refer to the model obtained from this step as the preliminary final model, which contains all the base questions in the questionnaire, except *Weight* and *Height*.

Before concluding any final model, we need to check for the goodness of fit of the model, and also for its prediction power. This problem will be addressed in the following section.

4.6 Evaluation of the regression model

Following the previous step, we need to check for the effectiveness of the model. What do we mean by the effectiveness? There are two things to be evaluated:

1. The goodness of fit.
2. The quality of prediction.

4.6.1 Goodness of fit

We denote the real value of the response variable by $y' = (y_1, y_2, ..., y_n)$, and the predicted value given by the model (*fitted value*) by $\hat{y}' = (\hat{y}_1, \hat{y}_2, ..., \hat{y}_n)$, the p independent variables are $x' = (x_1, x_2, ..., x_p)$. We say that the model fits well if:

- The difference between y and $\hat{y}'$ is small.
- The contribution of each pair $(y_i, \hat{y}_i)$ to that sum is unsystematic and small relative to the error structure of the model.

Pearson Chi-square Statistic and Deviance

First, let us denote:

- J: number of patterns/combinations of x' observed. For instance, when x' consists of 2 variables *sex* and *smoking status*, there could be possibly 4 patterns observed x' = (male, smoke), x' = (male, non-smoke), x' = (female, smoke), x' = (female, non-smoke). Let's remark that all subjects have different patterns, then $J = n$, if some of the observations have the same patterns, then $J \neq n$.
- $(m_j)_{j=(1,2,\ldots,J)}$: number of observations having the same pattern j. It follows that: $\sum_j m_j = n$
- y_j number of positive values (y=1) inside the j^{th} group. It follows that: $\sum_j y_j = n$

Unlike the linear regression, where the calculation of the distance between observed values and fitted values is obvious, there are several manners to measure this quantity in logistic regression. Recall that the logistic regression gives us the estimated probability for each covariate pattern, based on the value estimated, we can calculate $\hat{y}_j$, the fitted value estimating the number of positive cases in the j^{th} group as:

$$\hat{y}_j = m_j \hat{\pi}_j = m_j \cdot \frac{e^{\hat{g}(x_j)}}{1 + e^{\hat{g}(x_j)}} \tag{4.28}$$

where $\hat{g}(x_j)$ is the estimated logit.
The *Pearson residual* is defined as below:

$$r\left(y_j, \hat{\pi}_j\right) = \frac{\left(y_j - m_j \hat{\pi}_j\right)}{\sqrt{m_j \hat{\pi}_j (1 - \hat{\pi}_j)}} \tag{4.29}$$

$$= \frac{\left(y_j - \hat{y}_j\right)}{\sqrt{m_j \hat{\pi}_j (1 - \hat{\pi}_j)}} \tag{4.30}$$

$$\Rightarrow \chi^2 = \sum_{j=1}^{J} r\left(y_j, \hat{\pi}_j\right)^2 \tag{4.31}$$

The *Deviance residual* is defined as:

$$d\left(y_j, \hat{\pi}_j\right) = \frac{y_j - m_j \hat{\pi}_j}{\left|y_j - m_j \hat{\pi}_j\right|} \left\{ 2 \left[y_j \ln\left(\frac{y_j}{m_j \hat{\pi}_j}\right) + (m_j - y_j) \ln\left(\frac{m_j - y_j}{m_j (1 - \hat{\pi}_j)}\right) \right] \right\}^{1/2} \tag{4.32}$$

When $J = n$, the Deviance residual above is equal to the quantity in the equation (4.10). The statistic based on deviance is defined as:

$$D = \sum_{j=1}^{J} d\left(y_j, \hat{\pi}_j\right)^2 \tag{4.33}$$

According to [8], under the null hypothesis that the fitted model is correct, the distribution of χ^2 and d follows the chi-square distribution with $J - (p + 1)$ degree of freedom.

Reliability diagram

Because the logistic regression produces good approximations of $\pi(x_i)$, the idea here is to compare the estimated probabilities with the observed probabilities. The reliability diagram can be constructed afterwards based on that.

In the simplest setting, we suppose that $J = n$, in order to build the reliability diagram, we need to:

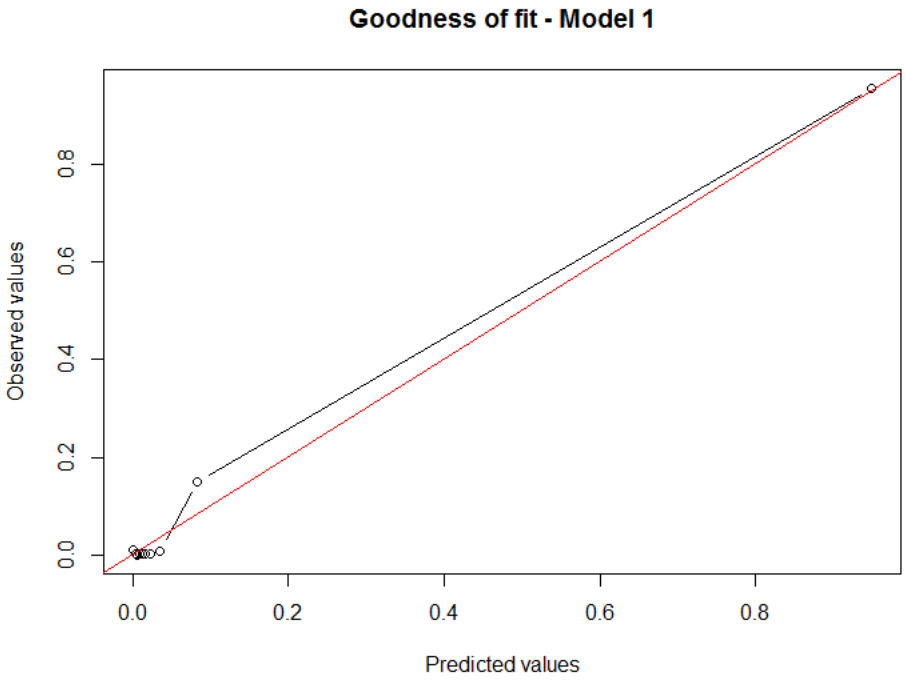
1. Sort the subjects by the estimated probabilities.
2. Let g be a number of group (usually $g = 10$). We regroup the previous subjects into g groups so that the first group contains $n'_1 = \frac{n}{g}$ subjects having the smallest estimated probabilities, ..., and the last group contains $n'_{10} = \frac{n}{g}$ subjects having the largest estimated probabilities.
3. In each group, calculate the proportion of positive outcome ($y = 1$)
4. At the same time, calculate the average of estimated probabilities.
5. Plot the diagram comparing the two quantities obtained from the previous step.

The result of the previous process is presented in the following table, where we choose $g = 10$:

Group number	Observed value	Predicted value
1	0,0111111111	0,0009602328
2	0,0017777778	0,0031308677
3	0,0008888889	0,0055433440
4	0,0026666667	0,0083039908
5	0,0017777778	0,0116077820
6	0,0026666667	0,0159080999
7	0,0022222222	0,0225002700
8	0,0084444444	0,0351187355
9	0,1502222222	0,0834243395
10	0,9542222222	0,9490957568

Table 4.5: Hosmer-Lemeshow test

1. Create a $2 \times n$ table where the rows correspond to the values of the response variable (i.e, $y = 0$, $y = 1$) and the columns correspond to J (n) different covariate patterns. Once the table built, we need to sort the columns based on the value of estimated probabilities $\hat{\pi}_j$, thus the first column of the table contains the smallest estimated values, and vice versa for the last column.
2. Let g be a number of group (usually $g = 10$). We regroup the previous pattern group into g group so that the first group contains $n'_1 = \frac{n}{g}$ subject having the smallest estimated probabilities, ..., and the last group contains the $n'_{10} = \frac{n}{g}$ subject having the largest estimated probabilities.
3. For $y = 1$ row, we calculate the predicted value by taking the sum of all the estimated probabilities over all subjects in a group.



The Hosmer - Lemeshow goodness-of-fit is defined as:

$$\hat{C} = \sum_{k=1}^{g} \frac{\left(o_k - n'_k \bar{\pi}_k\right)^2}{n'_k \bar{\pi}_k \left(1 - \bar{\pi}_k\right)} \tag{4.34}$$

where:

- $o_k = \sum_{j=1}^{c_k} y_j$, c_k is the number of patterns within the k^{th} pattern, hence o_k is the number of subject resulting to a response $y = 1$ inside the k^{th} group.
- n'_k is the total number of subjects in the k^{th} group.
- $\bar{\pi}_k = \sum_{j=1}^{c_k} \frac{m_j \hat{\pi}_k}{n'_k}$ is the average estimated probability.

Under the null hypothesis that the fitted logistic regression model is correct, $\hat{C}$ measuring the distance between the fitted values and the observed values follows the chi-square distribution with $g - 2$ degree of freedom.

Confusion table

As business required, calculating acceptance probability does not bring much value, so we need to deliver underwriting decision at the point of sales using fewer questions. To do that, we need to transform prediction probability into predicted underwriting decision, then we can check the

prediction quality by comparing predicted decisions with actual decisions in the most intuitive way. The transformation can be done via the following rule:

$$\hat{Y}_i = \begin{cases} 0 & \text{if } \hat{\pi}_i \leq \phi \\ 1 & \text{if } \hat{\pi}_i > \phi \end{cases}$$

The question then arises: what is the proper value of ϕ? Because the performance of our model is obviously very sensitive to the choice of ϕ, no matter how good the model is, if the value of ϕ is mistaken, good models can become very bad. The chosen value of ϕ is primordial and will be addressed later on. At this moment, suppose that we take the most common used value of ϕ, which equals to $0,5$.

Once predicted underwriting decisions are made, we can compare them to the observed value using *confusion table*. The result of the polychotomous logistic regression above is presented in the following table:

		AURA			
		1	2	3	4
Predictive UW	1	26895	165	1272	2
	2	0	405	44	16
	3	158	1	2215	26
	4	1	0	35	27
Number of cases		27054	571	3566	71
Accuracy		94,50%			
WPR		0,59%	29,07%	37,89%	61,97%

Figure 4.2: confusion table of polychotomous logistic regression

In the table above, we need to distinguish 3 different kinds of values:

1. **True / False positive**: the numbers at the diagonals of the table, and are the exact predictions of the model.
2. **False positive**: the numbers at the lower triangle of the table, and are in orange. Those values represent the wrong classifications of the model.
3. **False negative**: the numbers at the upper triangle of the table, and are in red. Those numbers represent the wrong classifications of the model.

False positive and *false negative* are both just wrong predictions, but we need to distinguish them because the consequences are very different:

- If we misclassify an application in the false positive part, it is referred when it should have been accepted, thus the underwriting process will require more human involvement, and consequently, increase circle time of underwriting. But in this case, underwriting model become more restrictive than automated underwriting.
- If we misclassify and application in false negative part, it is accepted when it should have been referred to underwriter for further examinations, underwriting circle gets better, but we take the bad risks into our portfolio, this causes damages to the business when the false negative applications are not managed in a proper way.

The following measures can be calculated from the confusion table:

- The overall *accuracy*, i.e, number of good predictions divided by the number of applications, gives us the general look about the classification quality of the model:

$$\text{accuracy} = \frac{26895 + 405 + 2215 + 27}{30062} = 94,5\% \tag{4.35}$$

- The *Wrong Prediction Ratios*, i.e, percentage of wrong predictions in each category, is equal to the sum of false positive and negative divided by the number of applications in each category.

$$\begin{aligned}
\text{WPR}_1 &= \frac{0 + 158 + 1}{27054} = 0,59\% \\
\text{WPR}_2 &= \frac{165 + 1 + 0}{571} = 29,07\% \\
\text{WPR}_3 &= \frac{1272 + 44 + 35}{3566} = 37,89\% \\
\text{WPR}_4 &= \frac{2 + 16 + 26}{71} = 61,97\%
\end{aligned}$$

Looking at the ratio calculated from the confusion table, we see that overall performance of the model is in general good with the accuracy = 94,5 %, especially for the cases in the first category with the wrong prediction only equal to 0,59 %. While the precision is not very satisfactory for the levels 2 and 3. For the last cases, rejected applications, our model gives really bad prediction, due to the lack of observations and unbalance proportion amongst all categories. Our observations are in-line with the point mentioned above that the classification is sometimes sensitive to the exposition of each levels and tends to favor the larger group where we have more observations over each pattern. The classification table is not a good measure for the fitness of the model, because we can easily imagine a situation where a logistic regression fits perfectly the observed data, but deliver the poor classification, for example, when the acceptance probability is approximately equal to 0,6. Then our classification has normally 40% of misclassifications. Even though classification table is not the good measure of the fitness, but it is used to measure the success of the model, because the goal of this study is not to fit the data, but to give good risk classification in order to increase customer experience. Good models are the models maximizing overall accuracy and minimizing WPRs. Note that in every classification, we have the trade-off between false positive part and false negative part as the accuracy is very sensitive to the arbitrate choice of ϕ, as different values of ϕ could favor one part and damage the other, and vice versa. Thus, the choice of ϕ depends on the risk appetite of each entities and on the market as well.

ROC,AUC

In case of dichotomous logistic regression, other ways of evaluating a model are available, the ROC curve (Receiver Operating Characteristic) and the Area Under the (ROC) Curve (AUC). These are just other ways of visualizing classification accuracy.

We dispose the following confusion table in the case of dichotomous logistic regression:

	0	1
0	a	b
1	c	d

The following measures can be calculated:

- Sensitivity $= \frac{a}{a+c}$
- Specificity $= \frac{d}{c+d}$

Contrary to the WPRs, the sensitivity and specificity measure the prediction accuracy in each category, the value of these measures depends strongly on the value of the chosen ϕ. If we chose the value for ϕ equal 0,5:

$$\hat{Y}_i = \begin{cases} 0 & \text{if } \hat{\pi}_i \leq \phi = 0,5 \\ 1 & \text{if } \hat{\pi}_i > \phi = 0,5 \end{cases}$$

the predictions will be:

		AURA	
		0	1
Predictive UW	0	27471	1347
	1	154	2290
Number of cases		27625	3637

Sensitivity	99,44%
Specificity	62,96%

Figure 4.3: Confusion table using $\phi = 0,5$

The value of $\phi = 0,5$ has some statistical advantages, but there are also inconvenience when the learning data have some special structures. The following table shows the predictions using a value of $\phi = 0,6$:

		AURA	
		0	1
Predictive UW	0	27497	1423
	1	128	2214
Number of cases		27625	3637

Sensitivity	99,54%
Specificity	60,87%

Figure 4.4: Confusion table using $\phi = 0,6$

By using different values of ϕ, the sensitivity increases by just 0,1 % with the trade-off of 2,09 % decrease of specificity. Depending on risk appetite of the entity, and the cost of each misclassification, one may say that 0,5 is better than 0,6 for ϕ. The same thing can be done for all possible cut points, the following table shows the values of specificity and sensitivity corresponding to the value of ϕ from 0 to 1, with increment of 0,1:

ϕ	sensitivity	specificity	$1 -$ specificity
0	0 %	100%	0%
0,1	54,61%	85,18%	14,82%
0,2	64,73%	77,45%	22,55%
0,3	71,47%	71,13%	28,87%
0,4	76,44%	66,65%	33,35%
0,5	80,26%	62,96%	37,04%
0,6	83,34%	60,87%	39,13%
0,7	85,88%	59,11%	40,89%
0,8	87,83%	57,68%	42,32%
0,9	89,46%	55,29%	44,71%
1	100%	0%	100%

The generalization of this table is presented in the graph below, it can be useful for choosing an a priori optimal ϕ. If our purpose is the classification, one might choose a ϕ maximizing both sensitivity and specificity, but this is impossible because a value of ϕ maximizing one will minimize the other. Instead of that, one can consider a value of ϕ maximizing the sum of these two measures.

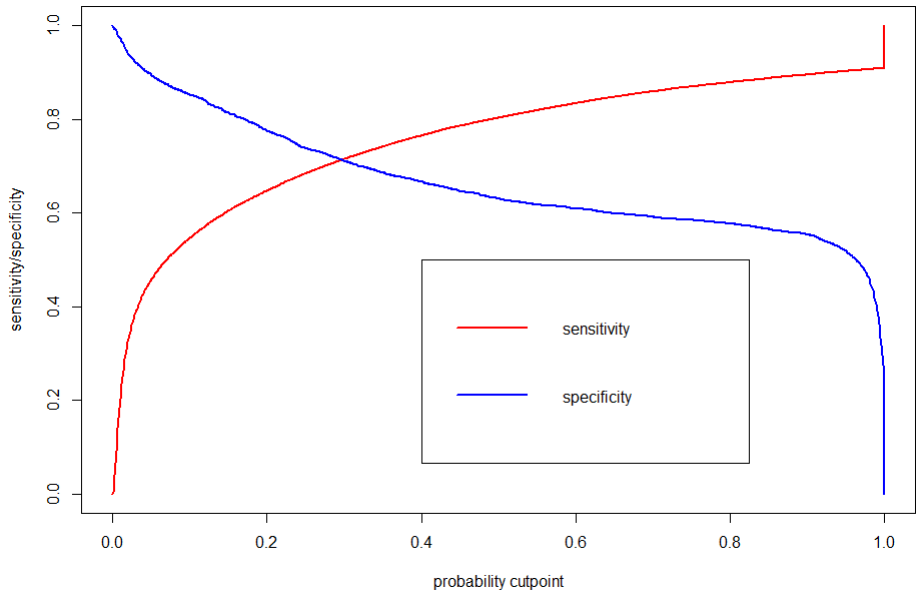


Figure 4.5: Sensitivity and Specificity for all possible cut points from 0 to 1.

The graph of sensitivity versus 1-specificity is the ROC curves shown below. The beginning point is always at (0,0) and the end point is always at (1,1). The worst model (or the model having no discrimination impact) gives the ROC curve in black. The perfect model (or the model predicting perfectly underwriting decision) gives the ROC curve in blue. Our actual model gives the ROC curve in red. The closer to the blue line is our ROC curve, the better our model is. Regardless the different impacts of false positive and false negative, the a priori optimal ϕ is the one corresponding to the point in the ROC curve the closest to the point (1,0) in the graph.

Base on the ROC curve, one can have a global view on the goodness of two models by comparing their Area Under the Curve (AUC). Our model has the AUC equal to 0,7542329. In the case of perfect discrimination, AUC ≈ 1, and contrary, in the case of no discrimination, AUC $= 0{,}5$. The higher the value of AUC is, the better our model is .

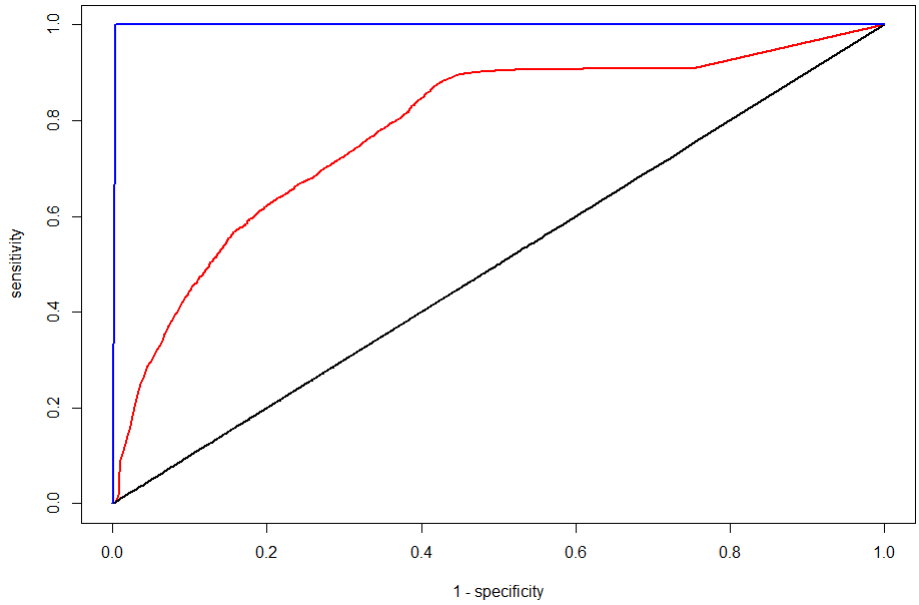


Figure 4.6: ROC curve

4.6.2 Evaluating the prediction power - the validation set approach

All the measures that we presented before are on the same data set, they give us the idea of how well our model fits the data, that doesn't mean that our model is good on new observations. In order to have good conclusions about the prediction power of our model, we need to assess the performance of the model on the new blind observations. If the model results in low test error rate, the use of such a model can be warrant in the practice.

This section mainly refers to the resampling methods in two kinds: Cross-Validation (Validation set, Leave One Out Cross-Validation, k-Fold Cross-Validation, ...) and the Bootstrap.

Validation set approach

This very simple approach consists of randomly dividing the data set into two sets: training set and testing set. The model is fitted on the training set, and the fitted model is used to predict the responses for the observations in the validation set. The resulting validation set error rate provides an estimate of the test error rate.

The disadvantage of this approach is that the validation error can be highly variable depending on the split between the training and the testing set. This variation can be limited if we have many data, and if the data are homogeneous.

Leave-One-Out Cross-Validation

This method borrows the idea of the validation set approach with a little refinement in order to limit its disadvantage. Instead of dividing the data set into training / testing set of comparable size, the testing set includes only one observation, and the rest goes to the training set. We repeat this on all observations of our data base. The model is trained on the training set, and evaluated on the test set N times. The average of the test errors can be used to estimate the MSE.

Comparing to the Validation Set approach, this estimation of MSE is less bias and more stable, but it is computationally expensive, especially when we have a complicated model or many data points, or both.

k-Fold Cross-Validation

One of the alternatives of the Leave One Out Cross Validation method is the k-fold Cross-Validation. Instead of choosing every observation as the testing set N times. We randomly divide the data set into training and testing set k time in such a way that the k testing set don't have a single common point and the union of the k testing sets makes up the entire data set. If $k = N$ the N-fold is equivalent to the Leave-One-Out Cross Validation approach. Usually we chose $k = 10$.

This method provides a proper estimation of MSE on the future data, less computational expensive than the Leave-One-Out approach, but its estimation is a little bit more variant.

4.7 Application (in underwriting and client scoring) to the AXA Italy portfolio

Here is the result of the logistic regression when applying to the portfolio of AXA MPS in Italy. In this model, we take into account 16 base questions, with the purpose of predicting the final underwriting decision taken by AURA.

We have in total of about more than 31 000 observations, we decided to separate randomly into 2 sets:

1. The training set, accounts for 80% of our total observations, is used to train the model.
2. The testing set, accounts for 20% of our total observations, the model built before on the training set will be tested on this testing set. Here we have more than 7 000 cases to test our model.

This is a relevant way to test the predictive power of the model, because we can use the model to predict the underwriting decision of the prospects that it never observed before. Then we can use the predicted underwriting decisions to compare with the observed underwriting decisions. The perfect model will give us the same predicted underwriting decisions as in AURA.

		AURA			
		1	2	3	4
Predictive UW	1	6082	30	248	2
	2	0	80	18	2
	3	36	0	488	6
	4	2	4	22	2
Number of cases		6120	114	776	12

Accuracy	94,73%			
FPR	0,62%	29,82%	37,11%	83,33%

Figure 4.7: Result of the logistic regression with 16 predictors

Let's note that logistic regression gives us only the probability of being accepted of a prospect. To have the predicted underwriting decision, we have to translate this probability to the number 0/1 (refer, accept). In this thesis, I used the rule below:

$$\text{Underwriting decision} = \begin{cases} 1 & \text{if } \mathbb{P}\left(Y = 1 \mid XX_1X_2...X_n\right) \geq \theta \\ 0 & \text{if } \mathbb{P}\left(Y = 1 \mid XX_1X_2...X_n\right) < \theta \end{cases}$$

The value of θ can be arbitrarily chosen from between 0 and 1. Usually we take the value of $0,5$ which is very intuitive.

This is the case for the binary dependent variable. If we have the dependent variable having more than two categories, we can take the underwriting decision as the most likely happen decision, i.e the category having the highest probability. By applying this rule, we obtain the *confusion table* 4.7.

The columns represent the underwriting decision given by AURA, i.e what we want to predict. The lines represent the predicted underwriting decision in the model taking into account all of 16 base questions. Let's us recall the coding of different decisions:

- 1: accepted cases with normal pricing
- 2: accepted cases with extra premium
- 3: cases referred to the underwriters for further examinations
- 4: declined cases

In total of 7022 testing cases, AURA accepts 6120 cases, accepts with extra premium 114 cases, refers 776 cases and declines 12 cases. Our predictive model with 16 predictors accepts 6362 cases, accepts with extra premium 100 cases, refers 530 cases, and declines 30 cases. In the confusion table, the good prediction lies in the diagonal, looking at the result, we can see that we delivered the good prediction for 6082/6120 accepted cases, 80/114 accepted cases with extra premium, 448/776

referred cases, and only 2/12 declined cases.

The *Accuracy* in the table represents the overall percentage of the good predictions, it is calculated by the sum of the numbers in the diagonal divided by the total number of the cases, here: Accuracy $= \frac{6082+80+488+2}{7022} = 94,73\%$ seems good.

It is also important to know the accuracy of our model with respect to each category (1, 2, 3, 4). The *FPR* (False positive rate) in the confusion table represents the percentage of the wrong predictions in each category. For instance, for the accepted cases: FPR $= \frac{0+36+2}{6120} = 0,62\%$. We can clearly see that our logistic regression works really well for the accepted cases, but not really well for the categories 2 and 3, and extremely bad for the declined cases. One of the reason is that in the accepted cases, we have the sufficient number of observations to train the model, as well as to test the model. This is not the case for the three others, especially the declined cases. Or the reason can be that logistic regression is not the good candidate for our problem.

In the confusion table, we need also to distinguish two kinds of wrong predictions:

1. Negative error: the numbers in red and situate at the upper triangle of the confusion table. For instance, the number 248 at the third column of the first line, these are the cases that our predictive model accepted but we should have referred to the underwriters. Those cases will reduce the cost paid to the underwriters (to review) but will be the problem for the risk management.

2. Positive error: the numbers in green and situate at the lower triangle of the confusion table. For instance, the number 36 at the first column and the third line, these are the cases that our predictive model referred but we should have accepted immediately. Those cases increase our underwriting circle time, increase the cost paid for the underwriters but will not be the problem of risk management.

The choice of θ affects highly the percentage of positive and negative value. An increase of θ decreases negative errors, but increases positive errors and vice versa. In our business, we prefer to have positive errors, rather than negative errors which will cost us very expensive in the future.

Chapter 5

Random Forest

As we see in the previous chapter, the logistic regression doesn't perform badly. But it remains inapplicable in the practice due to the high percentage of wrong predictions for the referred cases. As said at the beginning of the thesis, machine learning will bring the value to the business only if the right algorithm is chosen, so it's necessary to take into account at least another algorithm to test on our data, because of the time constraints, here we decide to take just one more algorithm named *random forest*, which is used as a benchmark of many data science competitions on Kaggle [1].

The idea of random forest lies in the decision trees algorithm, which is not a black-box, and is interpretable but results in many cases a very high variance. Random Forest algorithm was first introduced by Leo Breiman in 2001, combining Breiman's "bagging" idea and the random selection of features, introduced independently by Ho, Amit and Geman, with the aim of constructing a collection of decision trees with controlled variance. As logistic regression is an extended version of linear regression model, random forest is a generalized version of classification tree method by allowing multiple classification trees and averaging those results.

5.1 Tree-based methods

Tree-based methods are used to predict the value y_i from a vector $x_i \in \mathbb{R}^p$, it involves stratifying or segmenting the vector of predictors x_i into limited number of regions/ rectangles , and then fitting a simple model in each region in such a manner that the response variable takes more or less the same value in each region. Because the decision rules can be viewed as a tree, we call this kind of method the decision trees. The advantages of these methods are its simplicity and its ease to interpret. And the disadvantage of this method is its accuracy relatively comparing to other methods such as Bagging, Random Forest and Boosting which combine many trees with the intention of reducing the variance.

The rectangles can be obtained by successively dividing predictors $X_1, X_2, .., X_p$ into two separated intervals:

[1] the world's biggest platform organizing Data Science competitions on which companies and researchers post their data, then statisticians and data scientists from all over the world compete to produce the best models

$$\{X_i < t\} \text{ and } \{X_i \geq t\}$$

A common decision tree looks like this (here we took an example of decision tree on our data with only two predictors Age and BMI):

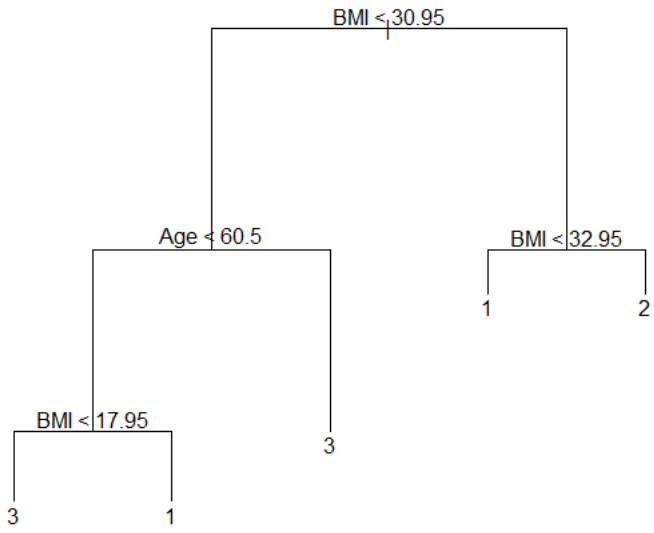


Figure 5.1: Classification tree predicting UW decision as the function of Age and BMI

In total, we have 5 different regions which are:

1. $BMI > 32.95$ with the predicted value 2 - accept with extra premium.
2. $30.95 < BMI \leq 32.95$ with the predicted value 1 - accept.
3. $BMI < 30.95$ and $Age > 60.5$ with the predicted value 3 - refer.
4. $BMI < 17.95$ and $Age < 60.5$ with the predicted value 3 - refer.
5. $17.95 \leq BMI < 30.95$ and $Age < 60.5$ with the predicted value 1 - accept.

The last division points of the tree are known as the *terminal nodes* or *leaves*, and the other points are *internal nodes*.

We can interpret the tree 5.1 as follow: the most uncertain cases (with respect to the final decision that can be conclude by AURA) are the ones having the $BMI < 17.95$ (can be interpreted as the malnutrition people) or having $Age > 60.5$ (the old people who commonly have health problems).

The most preferred cases are the young people ($Age < 60.5$) having BMI in between 30.95 and 32.95. If an individual has the $BMI \geq 32.95$, he (she) is still insurable but under certain conditions such as an application of the extra loading.

The predicted class c_j associated to the region R_j is just the most common occurring response within the region. But with tree-based methods, we can also estimate the probability of dropping in the class k of each individual, given the fact that the observation i lies in the region R_j, $\mathbb{P}\left(Y = k \mid X \in R_j\right)$ by the proportions of observations being of class k:

$$\hat{p}_k(R_j) = \frac{1}{n_j} \sum_{x_i \in R_j} \mathbb{1}_{(y_i = k)} \tag{5.1}$$

Hence:

$$c_j = \underset{k=1,\ldots,K}{\arg\max}(\hat{p}_k(R_j)) \tag{5.2}$$

where n_j is the number of observations in each region, and $\sum_j n_j = n$.

At this state, there is the presence of the trade-off between higher probability value in each region and the exposure. The less observations there are in the region, the higher chance the tree provides high estimated probability value. It is important to note that $\hat{p}_k(R_j)$ is not the true probability value, it is just an estimation, hence there will be high variation/ uncertainty if there are not enough observations in a region. This is the main weakness of tree-based methods.

Now two questions arise:

1. Choosing the split point t
2. The depth of the tree as there is the trade-off presented previously.

Theoretically we want to minimize the number of final misclassifications $\sum_i \sum_j \mathbb{1}_{(y_{i,j} \neq c_j)}$, but computationally it is unfeasible. Because, in a complete tree with at most M observations within each regions, meaning at least $\left\lfloor \frac{n}{M} \right\rfloor$ terminal nodes, i.e at least $\frac{\left\lfloor \frac{n}{M} \right\rfloor (1 + \left\lfloor \frac{n}{M} \right\rfloor)}{2}$ nodes. At each node, we could choose any of p predictors, meaning that the number of possibilities is:

$$p \cdot \frac{\left\lfloor \frac{n}{M} \right\rfloor (1 + \left\lfloor \frac{n}{M} \right\rfloor)}{2}$$

This is a huge number especially when we have an important number of observations. And we even haven't take into account the possible split points. Fortunately we have an alternative approach known as *recursive binary splitting*. Beginning at the top of the tree when all observations belong to a single region, successively splitting the feature space. At each step, the splitting point is chosen such as it results to the biggest drop in misclassification error, rather than looking further and find a set of splitting points minimizing the ultimate misclassification number which is evidently better but computationally unfeasible at this time. In other words, at each step, we need to select the predictor X_j and the cut point t splitting the feature space into two regions $R_1(j,t) = \{X_j < t\}$ and $R_2(j,t) = \{X_j \geq t\}$ with the goal of minimizing:

$$\sum_{i:x_i \in R_1(j,t)} \mathbb{1}_{(y_i \neq \hat{y}_{R_1})} + \sum_{i:x_i \in R_2(j,t)} \mathbb{1}_{(y_i \neq \hat{y}_{R_2})}$$

where $\hat{y}_{R_1}$ and $\hat{y}_{R_2}$ is the predicted class associated with the region R_1 and R_2.

Concerning the tree depth, the usual strategy is to grow a very deep tree (let's say each region has around 5 observations) in order to produce good predictions at the training set, but deep could easily lead to overfitting, thus we need to prune it at the end, meaning that we collapse some of its terminal nodes into the parent nodes. Smaller tree can not only reduce the variance, but also deliver better interpretation at the cost of a little bias. The preferred method for pruning a tree is *Cost Complexity Pruning*, because considering every possible sub-trees to find the one minimizing misclassification number is computationally expensive.

Let $|T|$ be the number of terminal nodes, we define the cost complexity criterion:

$$C_\alpha(T) = \sum_{m=1}^{|T|} \sum_{i:x_i \in R_j} \mathbb{1}_{(y_i \neq c_j)} + \alpha|T|$$

Base on the cost complexity criterion, for each value of α, there is an unique sub tree $T_\alpha$minimizing $C_\alpha(T)$. The tuning parameter α control the trade-off between tree complexity and its capacity to fit the data. The smaller the value of α is, the larger the tree is, and vice versa. $\alpha = 0$ results in the tree before pruning, and a very large value of α leads to a tree with 0 node. Now, we have a set of sub-trees T_α corresponding to a set of α. In order to choose the best tree from this set, we can perform k-fold cross validation to find the α (T_α) minimizing the test error.

5.2 Bagging and Random Forest

Bagging, Random Forest, Gradient Boosting are all developed from the basic idea of tree combining with bootstrapping technique.

5.2.1 The bootstrap aggregation or bagging

Tree-based methods can be a good candidate in fitting the training set, but it suffers from high variance in general even after pruning, meaning that the prediction error can be underestimated. Therefore, tree-based method is not efficient in practice.

The first simple way to reduce variance of tree was introduced by Breiman in 1994, the algorithm is called Bagging (**B**ootstrap **agg**regat**ing**). The way it works is very simple. We bootstrap a training set in order to produce many separate data sets, then perform tree-based methods on each of the bootstrapped set, and finally averaging the results enables us to obtain a low-variance model.

Let's say the number of bootstrapped training sets is B, and the predicted value of the individual i by the j^{th} tree is $\hat{f}^j(x_i)$, so the predicted value of the individual i is:

$$\arg\max_{k \leq K} \sum_{k \leq K} \sum_{b \leq B} \mathbb{1}_{\hat{f}^b(x_i)=k}$$

Note that each individual tree has low bias but high variance, by averaging many trees, the variance is reduced drastically at the cost of little bias. Baggings is proved to be very efficient in prediction problems, but the disadvantage of bagging is its interpretation. By averaging predicted values, it becomes impossible to have an intuitive and easily interpreted individual tree.

5.2.2 Random Forest

Even being built on different bootstrapped training set, individual trees in bagging are more or less correlated at some means. In more details, for example, let's say the variable X_j is strongly discriminant in a the whole training set, and the other variables are less discriminant. Hence, every individual tree in bagging will probably take X_j at the top split of the tree, and every tree will look similar and correlated. As the consequence, the variance of the learning model when the components are correlated is not reduced as much as in the situation of independent variables.
Breiman in 2001 introduced a new algorithm called Random Forest as an improvement over Bagging. The goal of Random Forest is to decorrelate individual trees. The only difference of Random Forest is just on the fact that at each split, only a random sample of features is considered, instead of the whole set of p predictors. In most of the cases, only $m = \sqrt{p}$ random features are taken into consideration at each split point, so even a weak predictor has a chance to be in the model, allowing random forest to capture very special signals in the data. When $m = p$, Random Forests works identically to Bagging.

Random Forest has two main tuning parameters:

- B: the number of trees.
- m: the number of random features to be selected at each split.

The value of these two tuning parameters could be chosen thanks to cross-validation. Usually, high number of trees doesn't make the model overfit the data, but it makes the model run slowly as the computer has to grow a large number of trees on different data sets. We can stop growing new trees when there is no reduction of the testing error. A small value of m can also help in the presence of a large number of correlated predictors.

In Random Forest / Bagging, it is not necessary to perform cross-validation in the purpose of tuning the parameters, because the test error is computed directly corresponding to each tree. Recall that each individual tree is grew based on the bootstrapped training set, which in general accounts for only two third of the training set, the rest can be used as the testing set to calculate the test error. Some statistical software produce Out-Of-Bag error straightforwardly used for parameter tuning.

As in bagging, the biggest disadvantage of random forest is the difficulty to interpret the results, comparing to tree-based methods, which are very intuitive and even are the better way than linear regression, which doesn't allow to interpret the results in some specific relations/interactions.

Implementation of Random Forest

Here are the results of a Random Forest on the Italy portfolio:

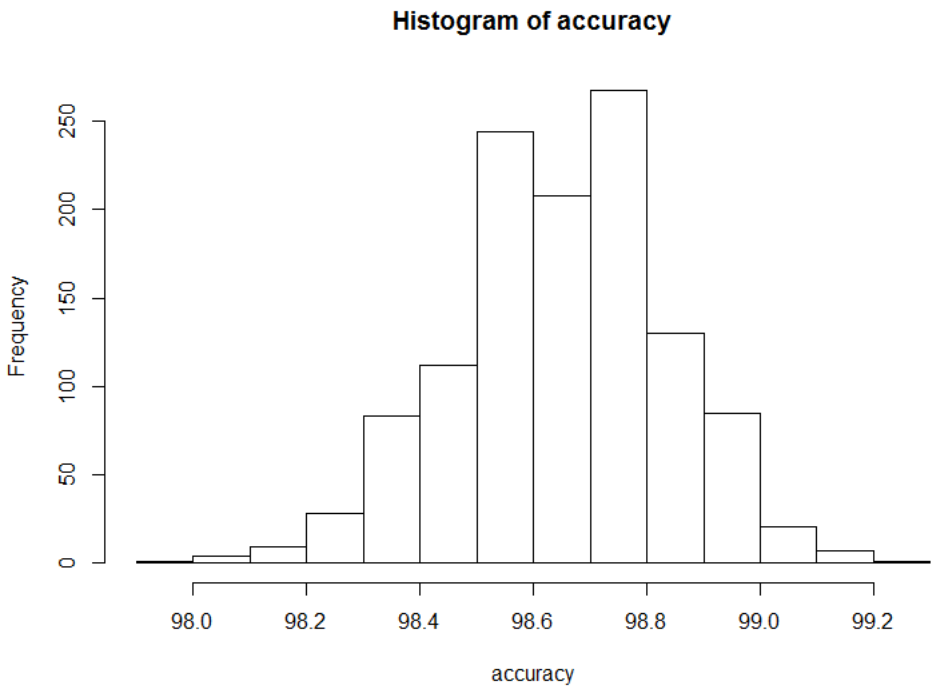
		AURA			
		1	2	3	4
Predictive UW	1	7945	0	20	4
	2	1	160	6	1
	3	43	2	1001	12
	4	0	0	2	5
Number of cases		7989	162	1029	22

Accuracy	99,01%			
WPR	0,55%	1,23%	2,72%	77,27%

We see clearly that Random Forest outperforms Logistic Regression in term of predictions. All the measurements (overall accuracy, WPR on each decision group) are better with Random Forest, especially the WPR within the second and the third group. Predictions stay at the bad quality for the forth group, as we don't have sufficient data points. Among 1029 referred prospects, Random Forest made good predictions for 1001 of them, and accept only 20 prospects (2,72%) as good risk. But the only thing we can say about these 20 cases is that Random Forest makes different decision compared to AURA, we don't know which is better between the two competitors, it can happen that for these 20 cases, Random Forest does the job even better than AURA, as they have more or less the same characteristic as other prospects. But at this point in time, we can not know more about the goodness of our model as we haven't had observations about the claims of these prospects yet. To evaluate the quality of Random Forest to AURA in terms of claims, we need to wait long time until see the actual claims of the prospects taken. But another strategy of comparison can be made, we can go deep into the process and see if these 20 cases "mis-taken" by Random Forest are all accepted by underwriters or not.

It is not fair to judge the performance of Random Forest based on only one random testing set, even if it is blind. It can happen that the good results observed above are just by chance as the results are at some sort of 'perfection'. All the measurements of goodness that we use until now depend on the way training and testing set split. If we change the training and testing set, all numbers change as well. So it is reasonable to question the stability of those measurements when varying training and testing set. By making large number of random splits between training and testing set, and by building for each split a model based on training set and predicting on unseen data, we can have a good idea of what could be the WPR when commercializing the algorithm.

Below is the distribution of overall accuracy, WPR within each sub-group obtained by performing 1500 independent random splits. A question of training time had arisen as it took 3 days of continuous training to get these graphs:



The model at the first sight is robust regardless the split between training and testing set. The overall accuracy is in most of the cases between 98,6% and 98,8%. In the best scenario, this could reach 99,2 %, and about 98% in the worst scenario. But our data are highly imbalanced, thus it is not wise to look only at the overall performance, we might be at the situation of overestimating the model. Bad surprises could probably occur when looking precisely into each underwriting's decision sub-group.

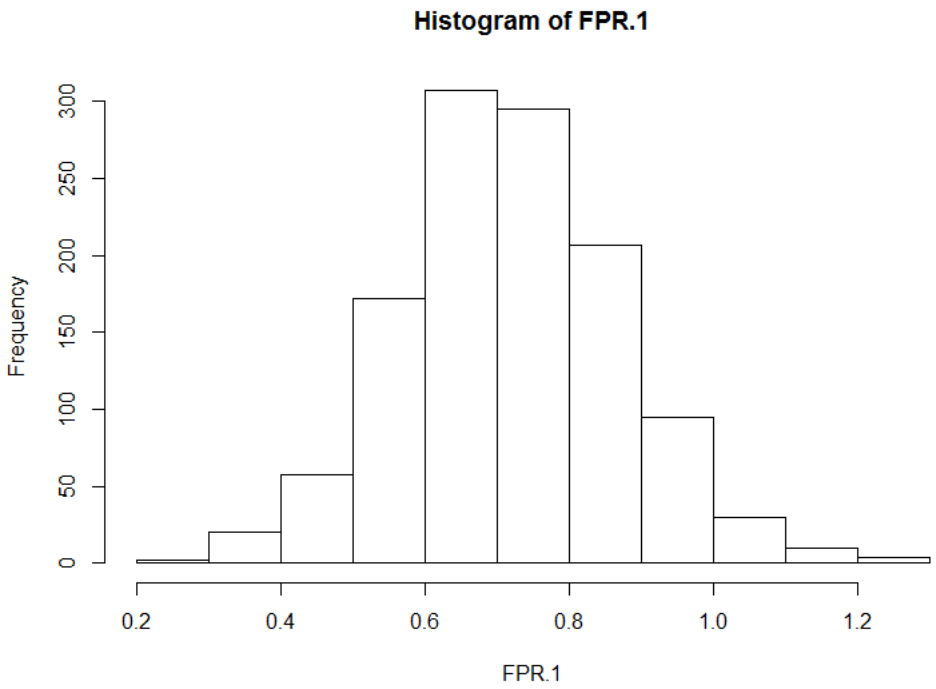
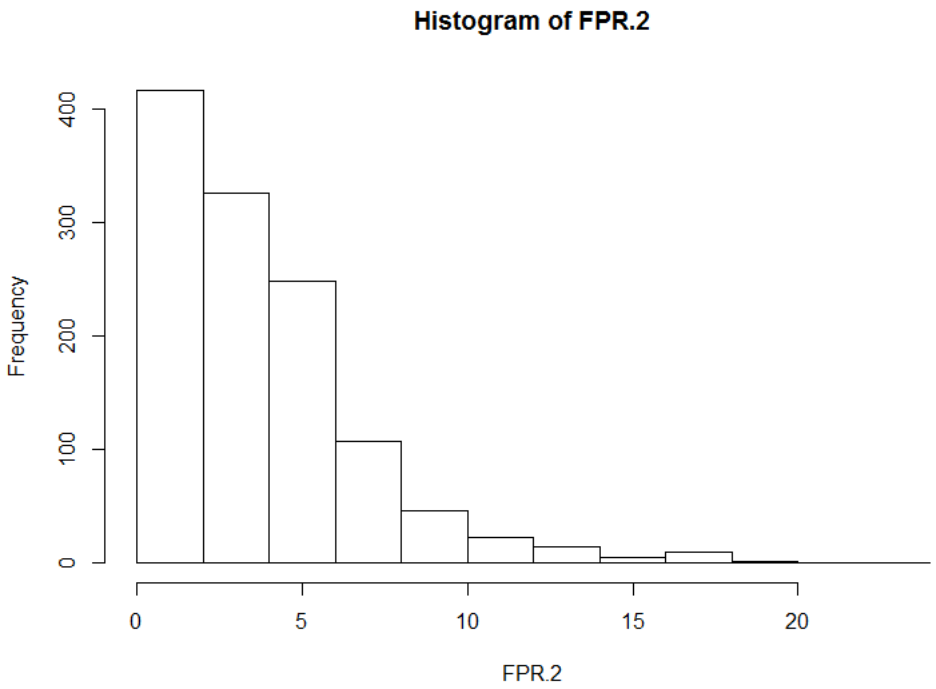


Figure 5.2: RF - Histogram of overall accuracy

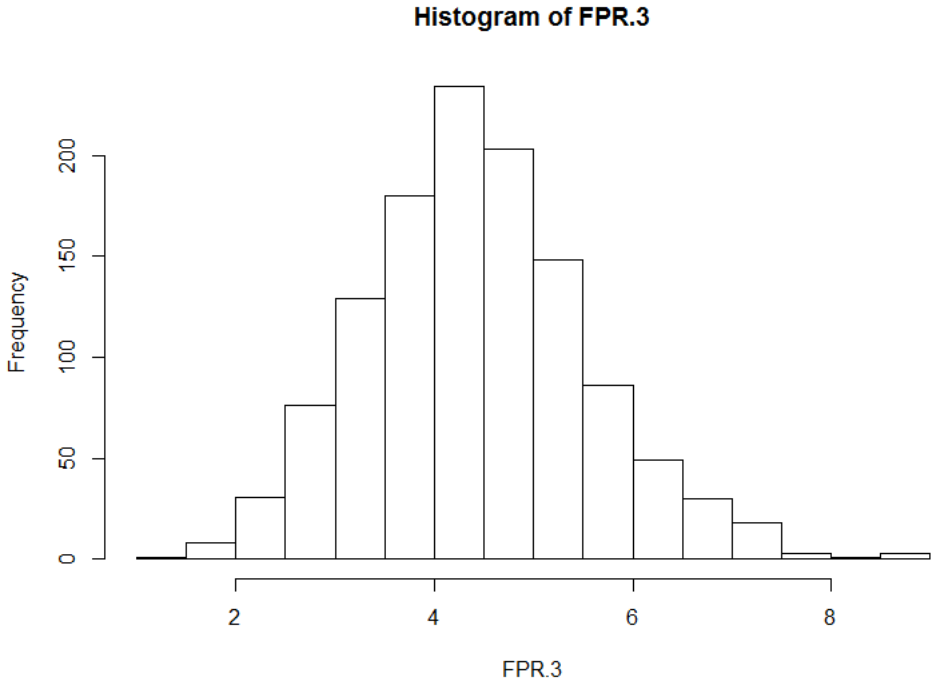
There is no surprise in the accepted group, which is understandable as the percentage of this group is very high in our data base, so the wrong path in this group of the tree is highly penalized. WPR distribution seems to be normal, with the mean of 0,7% and standard error only about 0,2%. In the extreme case, it could reach the value of more than 1%. In order to have better understanding of the model, one should also take a look at the measure of performance at the second and third group. In general, predictions will get worse because:

- There are less data points → robustness could be problematic.
- Their portion is very low relatively to the first group, thus Random Forest doesn't pay much attention on the wrong prediction of these groups if no particular treatment is made.

Below are the expected graphs:



The WPR in the second group (accept under certain conditions) has the higher variance than in the first group, which is not surprised as the data in this group are insufficient to make the model highly robust. But in the majority of the cases, our prediction error is below 5 % which is acceptable depending on the risk appetite of the entity.



Now we are seeing to the most important group of this study, the referred cases. The WPR distribution seems to be normal with the mean of 4-5% and the standard error of 2%. We make less than 1% prediction error in the best scenario and more than 8% in the worst scenario. The results are not as good as in the first group but still satisfying. One interesting point is that during the studying time, the WPR of each group progresses over time as we are getting more data points each month. With this reason, we believe that our model will even get better in the future as there are more data points for Random Forest to "learn".

What if we want to ask even fewer questions?

What is the benefit of reducing number of questions on the questionnaire?

- First of all, protection products are mainly sold through the bancassurance channel. AXA MPS, a joint venture between AXA and the oldest surviving bank in the world - Monte dei Paschi di Siena, is the perfect example where protection products are proposed by client's bank advisor. Fewer questions means less time spent by the advisor on proposing / underwriting the product, which is not their main jobs. Less time spending on this part will lead to significant higher performance of the advisor. In AURA's setting, we've seen some clients answering more than 200 questions, bank advisors in general don't like to work with this kind of situation. When the number of questions is limited, they are more likely to make efforts on proposing protection products.
- Secondly, fewer questions also lead to better customer experience being very important recently. Some questions related to family history are even very difficult to answer correctly, which could

become a barrier making a prospect having little need/interest of buying protection not to do so. And usually, those people are very profitable.

- Thirdly, we've experienced a significant number of abandonments during the process of answering the questions.
- Last but not least, it allows us to do a very interesting business that doesn't exist before. This point will be taken later on in this study.

Now what if we very ambitiously want to ask only 0 question, apart from the basic ones such as: the age, the sex, ... which in general is not intrusive to our clients?

		AURA			
		1	2	3	4
Predictive UW	1	8071	0	522	7
	2	0	164	19	1
	3	44	8	531	3
	4	0	1	0	8
Number of cases		8115	173	1072	19

Accuracy	93,55%			
WPR	0,54%	5,20%	50,47%	57,89%

Figure 5.3: Predictions when asking 0 question

Although the WPR within the first category remains at the low level, the overall accuracy is only 93,55%, which implies that we are giving bad predictions to about 7% of the prospects. Given that the rejection/ referred rate is about 11%. This ratio is better than random selection yet any bad selection will result in a bad product margin/ even losses. We don't want this solution at this point in time. But what about asking only 1 question?

Now the task is to see if the model with only one question performs well. Intuitively, the first work to do is to find out which questions are relevant to ask. Fortunately, with Random Forest, it is possible and easy to obtain **variable importance** of each predictor in constructing overall model. This could be used as a good starting point to rank questions in the limitation constraints.

As discussed before, the main advantage of decision tree over Bagging / Random Forest is its ease of interpretation. In Random Forest, we don't have this level of interpretation but we can extract the contribution of each variable in improving model's accuracy and rank them in order to gain insights about which variable is more "important" than the others in the model.

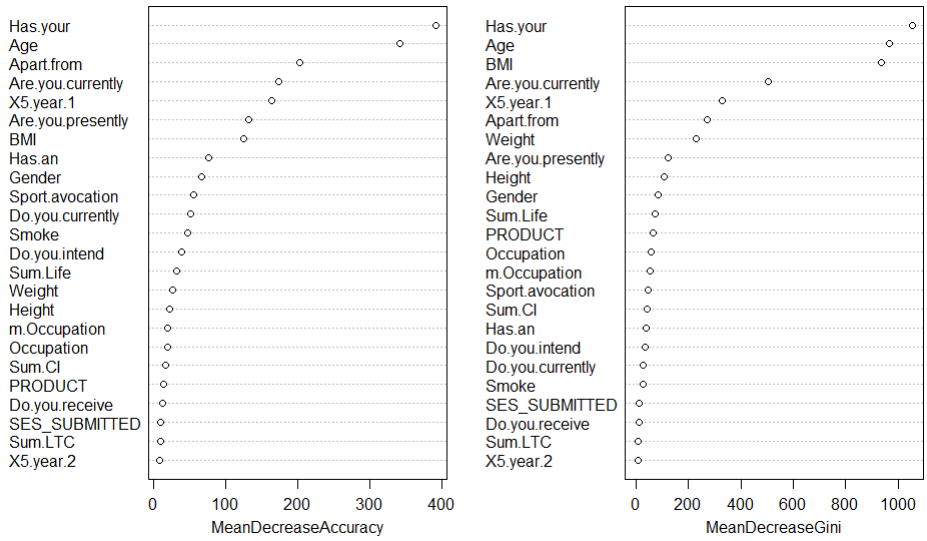


Figure 5.4: Importance level of each question using Random Forest, expressed relative to the maximum

The graph above represents:

- At the left hand side: *mean decrease in accuracy*, computed from permuting OOB data.
- At the right hand side: *mean decrease in node impurity* from splitting on the variable, averaged over all trees.

Apart from the basic questions, it is clear that the most important question is the one regarding prospects' family history. Now we can try to build a model with only basic questions plus information about the family history and see what happens:

		AURA			
		1	2	3	4
Predictive UW	1	8090	0	285	6
	2	0	167	13	1
	3	25	5	774	4
	4	0	1	0	8
Number of cases		8115	173	1072	19

Accuracy	96,37%			
WPR	0,31%	3,47%	27,80%	57,89%

Figure 5.5: Predictions when asking 1 question

The WPR at the first lot persists at low level. However, a significant improvement overall is observed. Model's accuracy improves from 93,55% to 96,37%. More important, WPR within the

second and third categories improves notably as well, with only one more question asked. More precisely, the wrong prediction rate at the third category (referred cases) is at 27,85%, which is relatively high. There are 285 cases wrongly accepted out of 1072 actually referred cases. Considering high sum insured in protection, we want to see what happen if we ask one / two supplementary questions.

		AURA			
		1	2	3	4
Predictive UW	1	8086	0	158	2
	2	0	171	5	0
	3	29	1	909	9
	4	0	1	0	8
Number of cases		8115	173	1072	19

Accuracy	97,81%			
WPR	0,36%	1,16%	15,21%	57,89%

Figure 5.6: Predictions when asking 2 questions

Not surprisingly, asking one more question improve considerably our predictions as the WPR within the third category is falling off twice (from 27,8% to 15,21%).

		AURA			
		1	2	3	4
Predictive UW	1	8055	0	105	2
	2	0	171	4	1
	3	60	2	963	9
	4	0	0	0	7
Number of cases		8115	173	1072	19

Accuracy	98,05%			
WPR	0,74%	1,16%	10,17%	63,16%

Figure 5.7: Predictions when asking 3 questions

Passing from two to three questions adds undeniable value to our predictions. But comparing to asking the second question, the value added of the third question is not at the same level.

Below is the recap of model's accuracy, as well as WPR over all categories of Random Forest's performance when asking from 0 to 10 questions:

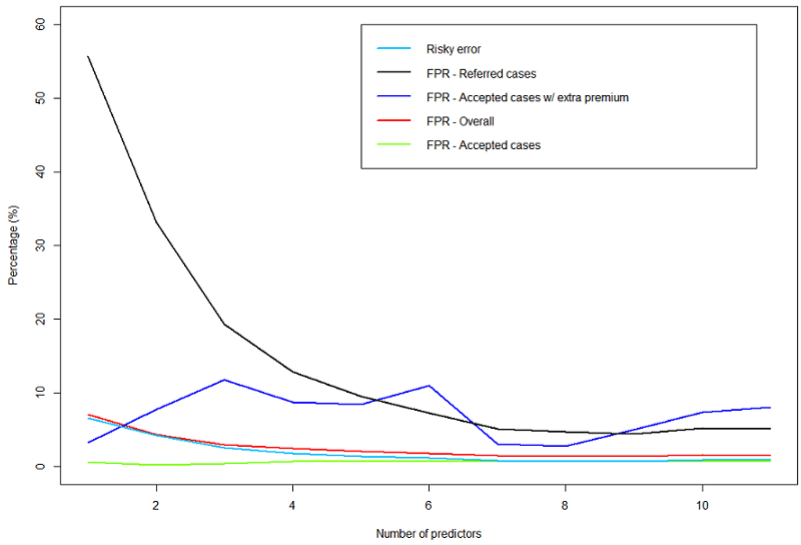


Figure 5.8: Prediction quality as a function of number of questions asked

The red line represents the overall prediction error. This line follows a downward trend as more information is added, which is intuitive but the trend's slop is closed to 0 starting from the forth question. The cyan line represents the WPR of the first category, this line is very flat due to the important percentage of the accepted cases. The blue line represents the WPR over the second category, no particular path is observed due to low number of observations, but its levels remains below 10%, which is a good sign. We've added also the arctic line representing the "risky error", i.e the percentage of red number in the confusion number above. We call them risky because they correspond to the accepted cases that should have been referred / rejected due to high risk. Allowing these cases to enter our portfolio will result in worse margin rate. We remark that the trend of the arctic line is highly similar to the red line's trend because the wrong predicted cases at the third lot are the main contributor of the overall errors.

The graph above furnishes an informative summary about the trade-off between model's quality and the number of questions asked. Asking ore questions means more information about the prospects, which leads to better predictions. But the value added starting from the seventh questions is not considerably high. The choice of the number of questions being asked depends on the process-elasticity of the local market, and on the risk appetite of the particular entity.

Telemarketing

Protection products are sold mainly through traditional channel, such as bancassurance as the example of AXA MPS. They are not sold on the Internet because of:

- Complicated Underwriting Process
- Their own characteristic: people don't need individual protection as their daily needs such as food, shelter, transportation, energy, clothing, ... nor obliged purchased products such as home insurance, car insurance, ...

It is more on the side of insurers searching their clients. And because of the complicated process, we are not able to sell our products online. The idea of Telemarketing approach arisen when American Express (an American financial company, specialized in payment methods) want to sell AXA MPS's individual protection products to its clients, with the constraint of asking at maximum 1 question on the telephone. It is without any doubt very challenging for the traditional approach to propose an underwriting strategy with only one question. This made us think about using Predictive Underwriting as an alternative in order to tackle this problem in the most statistically effective way

It is appropriate to contact the right prospects at the right time moment. As the margin is in general high, one more client can generate a considerable amount of profit, and can also mutualize the risk within the portfolio. But what are the right prospects?

- The ones being ready to buy the products, in order to economize the contacting cost, and to not bother our partner's clients. To tackle this, a model to quantifying the propensity to buy need to be built, with the same approach of Predictive Underwriting. But instead of predicting Underwriting decision, we can predict whether a contacted prospect will buy the product or not. Once propensity to buy is calculated for each prospect. We should rank them and give priority to the most probable to buy prospects. This model is a very interesting but out of the scope of this paper.
- Secondly, what questions to ask under the control of the quantity? What is the probability of being accepted? In this section, we try to answer these questions by investigating Predictive Underwriting. We will also analyze the trade-off between number of prospects to contact and the level of Underwriting risk taken.

The idea behind the utilization of Predictive Underwriting is to estimate the probability of being accepted of each prospect.

The graph below shows the sorted acceptance probabilities of more than 8000 prospects:

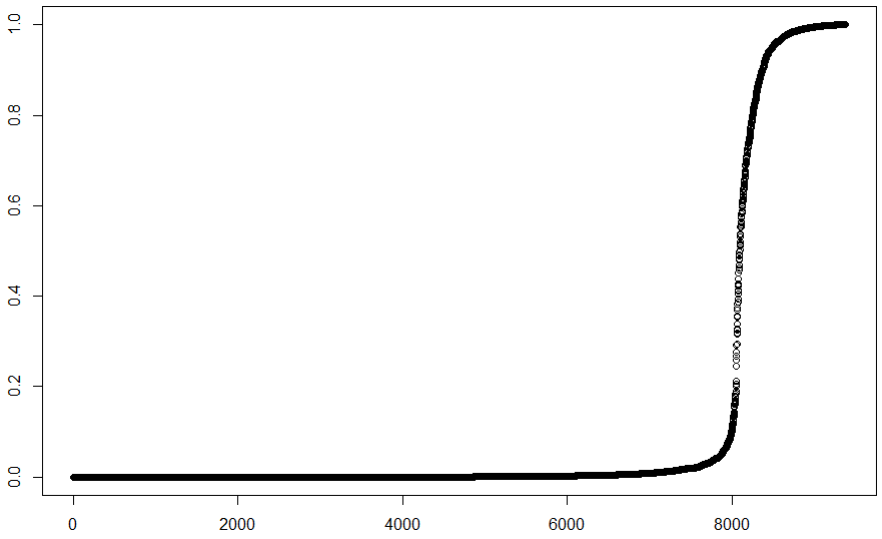


Figure 5.9: Output of the model with 10 questions

The x-axis represents the sorted prospects with their respective unacceptance probability represented by the y-axis. Because the probabilities are sorted, we see an increasing function (not absolute because some prospects have the same predicted probability). We remark that many prospects have approximately zero probability of being referred/ rejected (before the 6000^{th} prospect. We are very certain about accepting these prospects. Otherwise, the prospects lying at the bottom of the list (after the 8000^{th}) have very high probability, some even have 100% of chance of being referred/rejected if go through AURA.

Note that those are just probabilities of being not accepted. We can check the reliability of those predicted probabilities by seeing whether a prospect is accepted or not if going through AURA:

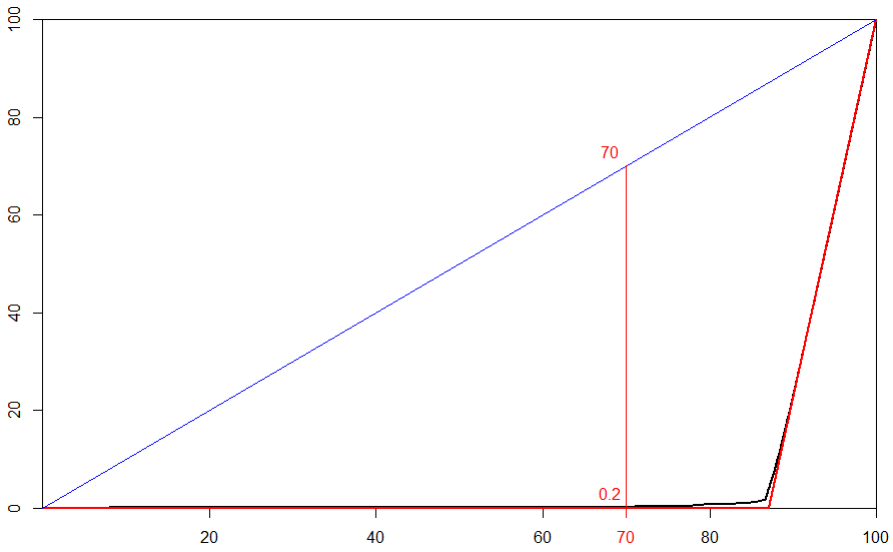


Figure 5.10: Back testing of model using 10 questions

How is this graph constructed? The black line represents our model. In other words, the x-axis represents the sorted prospects by the model, going from the left to the right, if the sorted prospect is actually not accepted by AURA, the black line goes upward by $\frac{100}{\text{total number of actually not accepted prospects}}$.

Thus the perfect model corresponds to the red line in the graph. The perfect model is capable of ranking perfectly more than 8000 prospects, i.e the 100% of the first ranked prospects before certain threshold are accepted, after that everyone is not accepted. The threshold here is the percentage of accepted prospects in the database (90%). The random model will be the one incapable of sorting prospects and is presented in the blue line. The closer to the red line the black line is, the better our model is.

A good strategy here for a very prudent entity, if we know in advance the information of these 10 questions, will be to select only the first 70% of prospects. This will result in approximately 0.2% of referred prospects entering the insured portfolio. Meanwhile random selection results in taking 70% of prospects who shouldn't have been accepted.

Note that in a Bayesian setting, probabilities are conditioned to the data. Here we use the model

taking into account 10 questions, giving us the black line almost identical to the red line. It is good, but not very interesting from the business point of view.

Come back to the use case of AMEX where the limited question number is one. The most relevant question remains family history as concluded anteriorly. The model gives us:

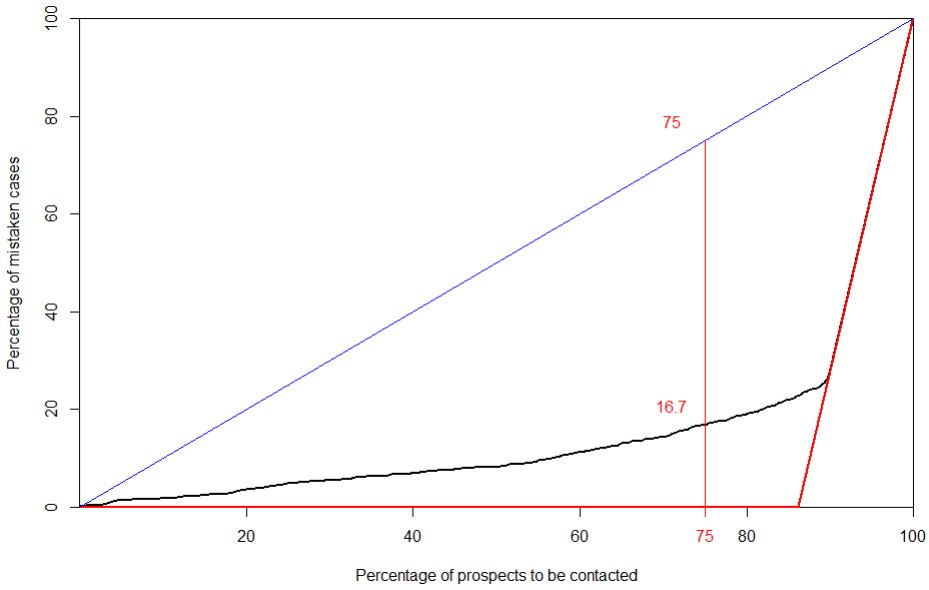


Figure 5.11: result of the model with 1 question

Given the list of 9000 prospects, approximately 10% of them should have been actually rejected, if we take the first 75% ranked prospects, there will be 16.7 % of not directly accepted prospects entering the portfolio, while the random selection will allow up to 75 % of not accepted prospects. The level of risk taken depends on the levels of the threshold chosen by the entity. The threshold 75% here corresponds to the probability of less than 10%.

It is interesting to see what the model of 2, 3, ... questions gives us:

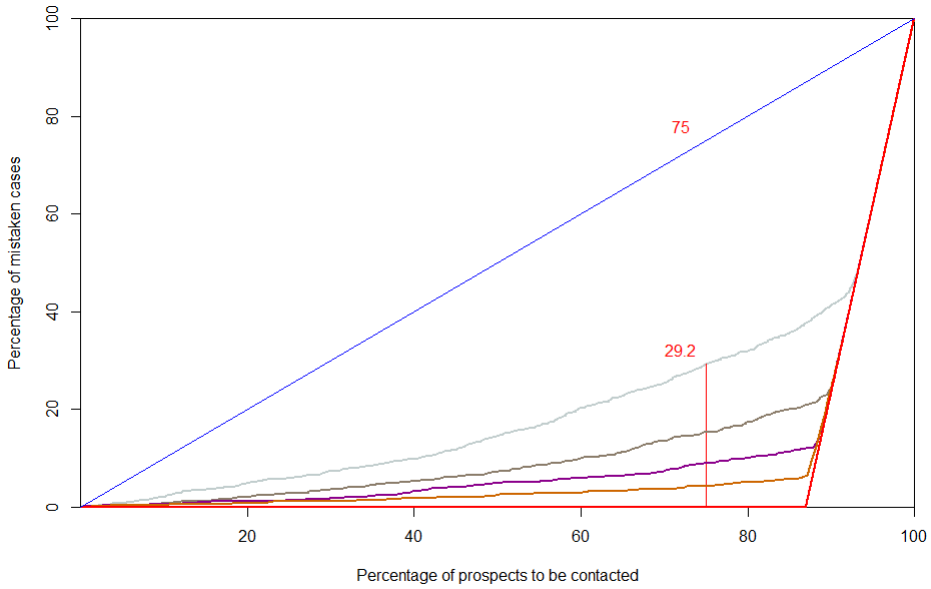
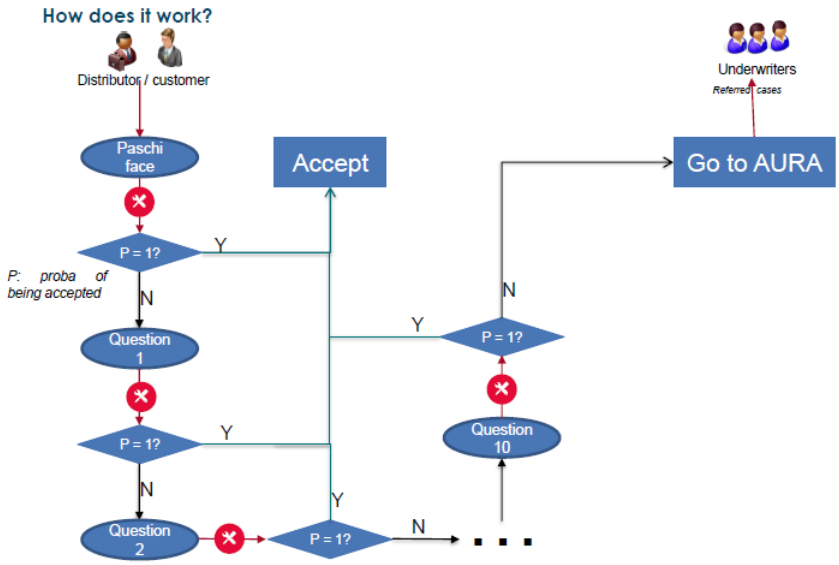


Figure 5.12: Telemarketing

From the upper line to the lower line are respectively the back-testing results of the model with 0 / 1 / 2 / 3 questions. We see clearly that the model gets better when adding a question. The "null" model using only basic information allows up to 29.2 % of not directly accepted cases entering the portfolio. The model with only three questions is already very good statistically.

Partial Predictive Underwriting

Random Forest gives us not only the prediction of underwriting decision, but also the probability of being accepted after answering each question. This characteristic of Random Forest makes us think about the idea of an iterative questionnaire as below:



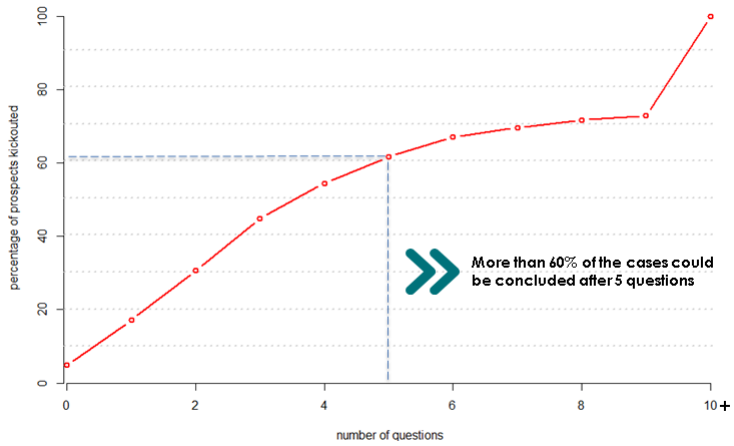
Given the basic information of a prospect (age, gender, ...) we make the first model predicting the probability of being accepted $\mathbb{P}_0$, if $\mathbb{P}_0 > \alpha$, he / she is accepted directly without asking any other questions.

If not, the first question will be asked, the choice of the first question is also from the variable importance of the Random Forest. Then after receiving the first answer, the second Random Forest will be run to calculate the probability of being accepted $\mathbb{P}_1$, again, if $\mathbb{P}_1 > \alpha$, the prospect will be accepted without asking the second question.

The process is going on until $\mathbb{P}_k > \alpha \quad (k \leq 10)$. If $\mathbb{P}_{10} \leq \alpha$, i.e the prospect gives already 10 most important answers to evaluating his / her health situation but our predictive model is still not sure about what decision to be taken, there are two possibilities, either we send the profile to AURA, in this case the prospect will answer more questions asked by AURA, and the decision should be correct, or we transform $\mathbb{P}_{10}$ into underwriting decision. The result of the second possibility which is quite good is already presented at the beginning of the study. But if we really want to assess carefully the prospect's health situation, the first option might be preferred.

α should be big enough in order to guarantee the quality of the taken decisions. But lower value of α could help us to reduce the number of questions. The choice of α is thus important but arbitrary.

We've tested the performance of our model on the extreme case ($\alpha = 1$), meaning that we are pretty sure about each decision taken. The result is quite satisfying:



- About 76% of the prospects answer less than 10 questions.
- The average number of questions asked reduces drastically to 5.5 %.
- While the quality of the decision taken is exceptionally good at 99.99 % overall accuracy.

Underwriting triage - What about the referred cases?

One of the reason why the underwriting process takes time is that the underwriters are working on the very complicated cases in which we request detailed medical tests. And they are money consuming and don't bring any value for AXA if the prospect is at the end rejected.

Data taken during several months of manual underwriting show that: amongst 1215 referred cases, only 371 cases are evaluated by our underwriters 3 months after finishing the questionnaire, 844 cases remain backlog. Instead of choosing randomly the cases to be evaluated amongst 1215 cases, we should focus on the cases having the highest chance of being accepted at the end. Predictive Underwriting can help underwriters increase the conversion rate at the manual underwriting stage by giving each application a probability of being accepted.

If we partition the probability space into three intervals $[0, 0.1), [0.1, 0.2), [0.2, 0.4), [0.4, 0.7), [0.7, 1]$, we can do a back-testing of the calculated probability against the actual decision taken as follow:

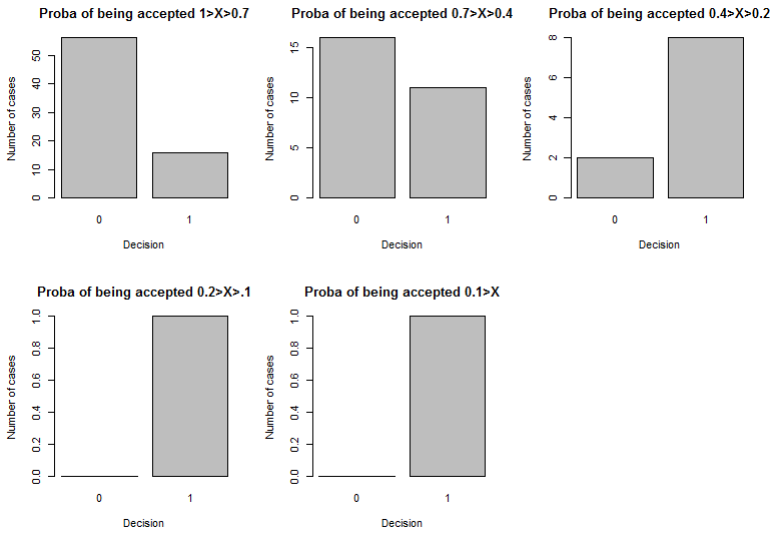


Figure 5.13: Back testing of calculated probability for manual underwriting (0 = accept, 1 = reject)

Giving this score to underwriters can help us distributing the available underwriters on the 'valuable' cases and thus increasing the conversion rate, and avoiding wasted time on the rejected cases.

5.3 Comparison of the results obtained from different methods

We can see clearly from the two models built previously that *random forest* outperforms *logistic regression* in term of prediction quality. Because random forest is able to capture very well non-linear (tree) structure that logistic regression doesn't. There exists other advanced Machine Learning techniques that are proven to be good in term of prediction such as Gradient Boosting, Support Vector Machine, ... But given the limited time and the accuracy of Random Forest, I did not spend additional time on testing other methods and decided to keep Random Forest as the final model for the Predictive Underwriting.

Chapter 6

Impact of Predictive Modeling

Commercializing Predictive Underwriting leads to better customer experience in term of optimizing underwriting process, which is beneficial to both clients and AXA. But it also allows risky prospects entering our insured portfolio. In order to mitigate this risk, we need to adjust our pricing strategy. The trade-off between the risk taken and customer experience is well detailed in previous sections. In this section, we take an example of a death benefit product pricing, how it is done in a traditional way and how to adjust that. This could serve as an indicator of how much risk to be taken, but it still depends on the elasticity of each market. The second point is important but is beyond the scope of this study.

In the first section, we don't going into the details on how to build a mortality table for the pricing. But at the second section, we will have to calculate the new mortality table of the new portfolio in the case of commercializing Predictive Underwriting. To do that, two kinds of mortality tables are necessary:

- The mortality table of the general population.
- The mortality table of the insured portfolio (after medical selection). This population is much less riskier than the general population. This mortality table is calculated by the actuarial team of AXA Italy.

6.1 Pricing of the life insurance policies

In this study, we will use profit testing to lead our pricing strategy. The idea behind is to fix the premium level in order to meet certain level of profitability. To measure the profitability of a contract, we need to consider a probabilistic perspective of every cash flows (in/out) generated. Two kinds of cash flows are generated during the contract life:

- Cash generated by the collected premium.
- Cash generated by the reserve.

Assumptions and notations

The purpose of a profit testing is to identify the profit that the insurer can get from the contract at the end of each time period, in this case at the end of each month, because the sum assured

will be paid eventually at the end of the month of death. To do this, the insurer needs to make assumptions about the expenses which will be incurred, the survival model for the policyholder, the rate of interest to be earned on cash flows within each time period before the profit is released and possibly other items such as an assessment of the probability that the policyholder surrenders the policy. We will go though each of these elements in details in this section.

- For the ease of presentation, we suppose no lapse during contract life.
- N: policy term.
- SA: sum assured.
- x: entry age of the insured.
- i: interest rate, in this study we make a hypothesis that i is unchanged during contract life.
- $(C_t)_{t=1,\dots,N}$: commission at each premium collection. It represents the acquisition cost of the policy paid by the insurer. The amount of C_t is usually equal to a portion of the premium collected at each time.
- Initial expenses: is the fixed amount paid at the acquisition of the contract. It is paid only one time at $t = 1$, and it is practically composed of 3 components: the first one relates to the SA, the second one relates to the first premium, and the last component is a fixed amount.
- Renewal expenses: usually equal to a portion of the premium acquired at time t.

Different types of cash flows generated during the contract life

Now we have all the necessary elements to begin our pricing process. Let's first take a look at different types of cash generated by the policy:

$$GCF_t = \sum_j (P_j - \text{Commission}_j - \text{Initial Expeses}_j - \text{Renewal Expenses}_j) * (1 + i) - \mathbb{E}[S_i]$$

where:

- GCF_i: Gross Cash Flow at time i.
- P_j: premium collected at time j.
- $\mathbb{E}[S_i]$: expected Cost of Death Outgo, is calculated by: $\mathbb{E}[S_i] = SA * q_{x+i}$.
- q_{x+i}: probability of the death occurred during the next month for a person aged $x + i$. This will be the only element that changes in the context of Predictive Underwriting vs Traditional Underwriting.

Then we need to look at what are the types of cash flows generated by the **reserve**.

We can remark from the previous section that the GCF_t can be negative or positive depending on the moment we look at it. This occurs because the level of premium is more than sufficient to pay the renewal expenses and the expected death claims in the early years, but, with an increasing probability of death, is not sufficient in the later years. The reserve is the actual amount of money held by the insurer to meet future liabilities. It can be calculated separately.

Let's denote R_i the reserve amount at time i . We have the formula:

$$R_t = \frac{(-P_{t+1} + \text{Commission}_{t+1} + \text{Initial Expenses}_{t+1} + \text{Renewal Expenses}_{t+1} + \frac{\mathbb{E}[S_{t+1}] + R_{t+1} * (1 - q_{x+t+1})}{1+i})}{1 - q_{x+t}}$$

subject to $R_N = 0$ because no money needs no be set aside at the end of the contract. This formula can be reformulated as:

$$R_t * (1 - q_{x+t}) + P_{t+1} = \text{Commission}_{t+1} + \text{Initial Expenses}_{t+1} + \text{Renewal Expenses}_{t+1} + \frac{\mathbb{E}[S_{t+1}] + R_{t+1} * (1 - q_{x+t+1})}{1 + i}$$

The reserve of this year plus the premium received the next year need to be adequate to pay all the cash outflows (commissions, expenses, expected claim) and also the reserve of the next year in the probabilistic perspective where comes the term $(1 - q_{x+t})$.

Until now we can calculate the Profit (before solvency) at each time period of the contract as follow:

$$\text{P1}_t = \Big(GCF_t - R_t + \underbrace{\underbrace{(R_t - R_{t-1})}_{\text{Increase in Reserve}} * (1 + i)}_{\text{Investment income on reserve}} \Big) * (1 - \text{tax})$$

The profit at each time will be equal to the Gross Cash Flow (cash generated from the policy) minus the Reserve that the insurer needs to set aside, plus the investment income on that money set aside and minus tax.

In every insurance company, for each client, we need to set a capital equal to the Solvency requirement. As the calculation of solvency capital is not the center of this study, for the simplicity, we will use the Solvency I convention in our calculation.

$$\text{RSC}_t = R_t * k_1 * 1.5 + (SA - R_t) * k_2 * 1.5$$

where RSC_t represents the Required Solvency Capital at t, and k_1, k_2 factors for Solvency margin requirement, its value depends on the market.

The ultimate profit of a contract at time t (after tax, after solvency requirement) is:

$$\text{Profit}_t = \text{P1}_t - \text{RSC}_t - \text{RSC}_{t-1} + \text{RSC}_{t-1} * (1 + i)$$

Until here we have sufficient elements to calculate the Present Value of Future Profit:

$$PVFP = \sum_{t=1}^{N} \text{Profit}_t * (1 - q_{x+t-1}) * \frac{1}{(1+i)^t}$$

And the Present value of future Premium can be calculated following the same principle:

$$PVP = \sum_{t=1}^{N} P_t * (1 - q_{x+t-1} * \frac{1}{(1+i)^t}$$

And finally, the profit margin of the contract is:

$$PM = \frac{PVFP}{PVP}$$

You can already remark that, given the mortality table giving the value of q_x the Profit Margin of the contract is a convex function of premium. In this pricing theory, the goal of setting premium is to meet certain level of profit margin, so one can easily use iterative resolution to find a level of appropriate premium. This value is unique for each targeted profit margin because it is a convex function of premium.

6.2 Pricing taking into account the prediction errors

One of the most important elements of Pricing is the mortality table. Predictive Underwriting will allow entering of bad risks into our portfolio, this effect will have negative impact on the mortality table of the insured portfolio. The more risks we take in Predictive Underwriting, the bigger the negative effect on mortality table will be. This will result in a price increase. The question here is how much should we increase the price in order to mitigate the negative impact of Predictive Underwriting on our insured portfolio.

So the question that we need to answer now is: what will be the new mortality table after applying Predictive Underwriting?

There will be two groups of people in our new portfolio:

- Group 1: people accepted by Predictive Underwriting and should be accepted by underwriters / AURA.
- Group 2: people accepted by Predictive Underwriting and should not be accepted by underwriters / AURA. (prediction's errors)

To calculate the new mortality table, we need three things:

1. The mortality table of the group 1. We can use the actual mortality table of the insured portfolio for this group. In this case, we get access to the mortality table of the death benefit product built by AXA MPS's actuaries.
2. The mortality table of the group 2.
3. The proportion of people between two groups. This can be estimated from the Random Forest model.

We've never observed the mortality table of the second group, so it is impossible to have their "correct" mortality table. But we can estimate it by comparing the mortality table of the insured portfolio to that of the general population. So, let's denote:

- A: number of the accepted prospects.
- R: number of the refused prospects.
- q_x^a: the mortality model of the accepted portfolio.

- q_x^r: the mortality model of the refused people.
- q_x^{ar}: the mortality model of the general population.

The total number of accepted and refused people must be equal to the size of the general population at the beginning and at the end of each year, hence:

$$\begin{aligned} A * q_x^a + R * q_x^r &= (A + R) * q_x^{ar} \\ \rightarrow q_x^r &= \frac{(A + R) * q_x^{ar} - A * q_x^a}{R} \end{aligned}$$

q_x^a is the mortality table calculated by AXA MPS's actuaries. A, R can be estimated from the AURA's population. This approximation is not perfect because people wanting a life insurance contract will have different death probability from the general death probability. The best approach is to compare the insured population filtered by the underwriting process to the insured population not filtered by any medical selection process. But in Italy we don't have any product not having medical selection, so this solution is not possible.

Now that we estimated q_x^r, we can combine it with the q_x^a using the expected prediction error. This prediction error can be estimated on the one single testing set, but a better way is to randomly divide the training / testing set many times, training the model on the training set, test it on the testing set. Then we can have the distribution of the prediction error on each class. There could be 3 scenarios:

- Normal scenario: the prediction error for each class is the mean of the distribution.
- Optimistic scenario: the prediction error is the 10% quantile.
- Pessimistic scenario: the prediction error is the 90% quantile.

In order to simplify the presentation, let's suppose that the results of the Predictive Underwriting is as follow:

Results		Observation	
		Accept	Referred
Prediction	Accept	a	c
	Referred	b	d

In which:

- a will be directly accepted into the portfolio.
- b will be also accepted after being referred to the underwriters.
- c will be partially mistakenly accepted by the Predictive Underwriting.
- d will be examined and certain client will be accepted by underwriters.

According to the historical data, about $x = 30\%$ referred prospects are rejected by underwriters after examining their medical information. Using this as a hypothesis, the portfolio size of Predictive Underwriting will be:

$$a + b + c + d * (1 - x)$$

Hence, if we note q_x^n the mortality model of the new population:

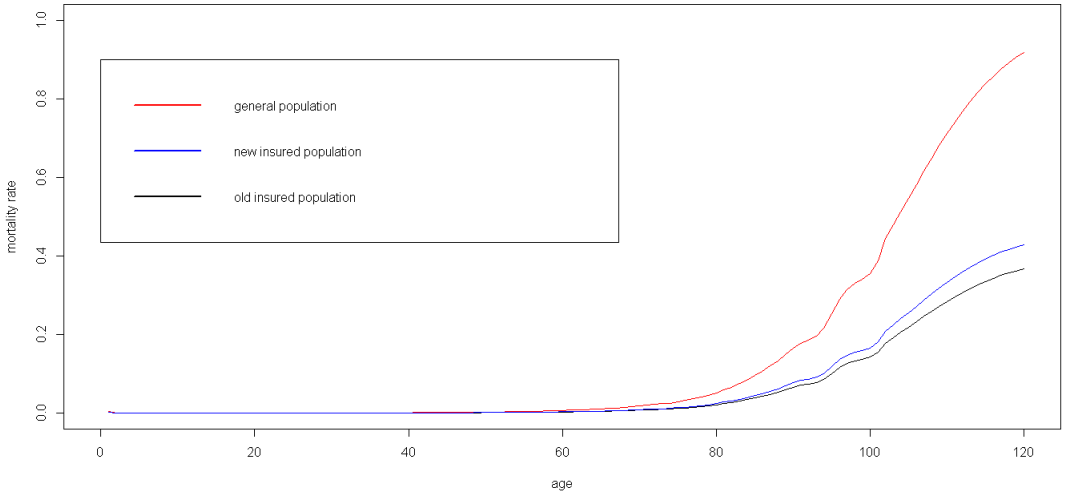
$$\underbrace{\big[a + b + c + d * (1 - x)\big] * q_x^n}_{\text{new population}} = \underbrace{\big[a + c * (1 - x) + d * (1 - x) + b\big] * q_x^a}_{\text{correctly accepted prospects}} + \underbrace{c * x * q_x^r}_{\text{prediction error}}$$

We have this formula because in c: only $x * c$ is mistakenly accepted by Predictive Underwriting, and the other $c * (1 - x)$ is correctly accepted, although they are the prediction errors.

So the final mortality table of the new population is:

$$q_x^n = \frac{\big[a + c * (1 - x) + d * (1 - x) + b\big] * q_x^a + c * x * q_x^r}{a + b + c + d * (1 - x)}$$

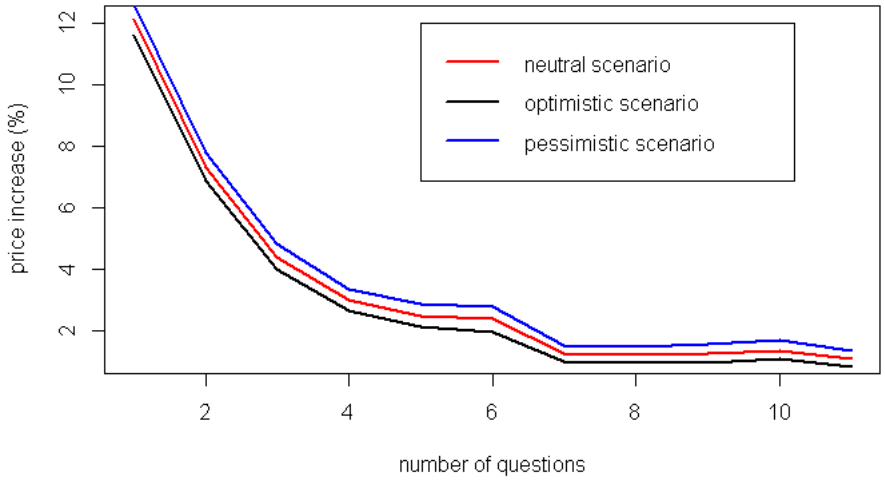
The graph below shows the new mortality table if we ask only 3 questions:



In this figure, we see the three mortality models of three different populations: the first one in red is without any medical selection, the second one in black corresponds to the population with traditional underwriting, the last one in blue is the mortality table of the predictive underwriting. We see clearly that for young people, there isn't much mortality difference, but from the age above 60, medical selection allows us to reduce the mortality rate by quite a lot. Applying Predictive Underwriting means allowing bad risks to enter our portfolio, hence the difference between the blue and black curve.

With this new mortality table calculated, we are able to play around with the risk premium pricing. The idea here is to calculate the trade-off between asking x question and adding $y\%$ premium as

the loading to mitigate the negative effect of Predictive Underwriting for a given client's profile, the results of a normal profile (40 year old, male, non - smoker) are shown in the graph below:



The difference between the three scenarios comes from the calculation of a, b, c and d. Firstly, the data set is 1000 times randomly separated into the training / testing set, for each separation, a random forest is built on the training set, and the performance is calculated on the testing set. At the end, we have four vectors: $A = [a_1, a_2, ...a_{1000}], B = [b_1, b_2, ..., b_{1000}], C = [c_1, c_2, ..., c_{1000}], D = [d_1, d_2, ..., d_{1000}]$

- The neutral scenario: $a, b, c, d = \mathbb{E}[A], \mathbb{E}[B], \mathbb{E}[C], \mathbb{E}[D]$.
- The optimistic scenario: a, d are the true negative and positive so we take the quantile at 90%, and vice versa for b, c (quantile at 10%):

$$a = Q_{0.9}[A]$$
$$b = Q_{0.1}[B]$$
$$c = Q_{0.1}[C]$$
$$d = Q_{0.9}[D]$$

- The pessimistic scenario:

$$a = Q_{0.9}[A]$$
$$b = Q_{0.1}[B]$$
$$c = Q_{0.1}[C]$$
$$d = Q_{0.9}[D]$$

This graph shows the trade-off between number of questions (aka underwriting efficiency) and premium loadings. Clearly, asking fewer questions leads to lower model's accuracy, meaning more bad risks into the portfolio. This results in high premium increase as shown in the graph. We can also see that, the first questions allow us to reduce a lot of premium loadings comparing to the last questions. We can also see that asking the $8^{\text{th}}, 9^{\text{th}}$, and 10^{th} questions doesn't bring much value comparing to asking only 7 questions.

This can be very useful for local actuaries / pricing analyst in order to decide the optimal number of questions depending on the characteristics of each market. If the local market has high price elasticity, then the solution might be to ask 7 questions as it would not affect very much the price of the product. In other case, the number of questions depends highly on the elasticity of the client on the price and on the underwriting process.

Chapter 7

Conclusion and Further Room for development

This thesis presented one of the first Data Science related works done within AXA in Life Insurance, an initiative regarding the improvement of the underwriting process made by the Individual Protection team of AXA Global Life. First of all, we saw how the underwriting process is done traditionally, then we saw how it was improved with recent technology development. And finally, how it will be innovated in the future with the help of Data Science, and a lot of data coming to the pocket of insurers.

The first part of the study showed the structure of the data, together with some descriptive analysis of the data.

In the second time, we investigated our work on the most common and basic technique named *logistic regression* in order to predict the underwriting decision by taking into account only 16 questions (in stead of 210 questions in one case of our portfolio). The overall accuracy of the model seems good, so is the accuracy of the accepted with standard cases. But the accuracy for the referred cases remains unacceptable (37,11 %). If we use logistic regression as the engine of our predictive underwriting, not only the we don't decrease very much the number of questions (because only 10% of the cases answer more than 16 questions), but also we take very high risk due to the low accuracy rate in the referred cases.

As the next steps of our work, we applied the second machine learning algorithm called *Random Forest*, which is used as the benchmark in many data science competitions on Kaggle, the world biggest platform organizing competitions for the data scientists. Random Forest outperforms Logistic Regression in every aspect of our study. The overall accuracy of the model is very satisfying, together with the accuracy on the sub-populations corresponding to each underwriting decision, except the directly rejected cases where we have very limited observation.

The satisfying results of Random Forest allow it to be very valuable for AXA's business in term of:

- Simplifying the underwriting process by firstly simplifying the questionnaire.
- Creating new business 'Telemarketing', in which only few questions are necessary to deliver the underwriting decision (in the expense of loadings on the premium). Combining with the

propensity model, this can be a profitable business for AXA.

- Scoring the referred prospects in order to help underwriters to work more efficiently.

We saw throughout the thesis that Data Science in general and Predictive Underwriting in particular will bring values for AXA's business in various way. But in practice, implementing Predictive Underwriting means that we will have less information about the prospects, so it can be difficult to know whether our model remains stable or not, perhaps the model performs well at the beginning, but it can be eventually deteriorated one year after the building time of the model. So we need to implement a process to control the performance of the model.

There are still rooms for development following this study. First of all, more external data, bank data should be integrated into the study. Concerning the modeling part, the first idea is to use another algorithm to challenge the retained model: Random Forest. After that, in the partial underwriting approach where the order of question asked is the same for everyone, we can try to make an iterative questionnaire, the following question depends on the response of the previous question. Another idea is that, instead of optimizing the confusion table, one possible alternative is to calculate the mortality rate associated to people in each sub-group of decision, then we can tune the parameters of the model in order to minimize the mortality of the insured portfolio, and so on.

List of Figures

2.1 The structure of AURA 17
2.2 Predictive Underwriting 18

3.1 Decision of AURA 22
3.2 Spine plot of BMI and Decision 23
3.3 Spine plot of Decision and Gender 25
3.4 Spine plot of BMI and Gender 26
3.5 Spine plot of Gender and Smoking status 27
3.6 Spine plot of Gender and age 28
3.7 Spine plot of Smoke and Decision 29
3.8 Spine plot of Age and Decision 30

4.1 Relationship between BMI and Weight 53
4.2 confusion table of polychotomous logistic regression 58
4.3 Confusion table using $\phi = 0,5$ 60
4.4 Confusion table using $\phi = 0,6$ 60
4.5 Sensitivity and Specificity for all possible cut points from 0 to 1. 61
4.6 ROC curve 62
4.7 Result of the logistic regression with 16 predictors 64

5.1 Classification tree predicting UW decision as the function of Age and BMI 67
5.2 RF - Histogram of overall accuracy 73
5.3 Predictions when asking 0 question 76
5.4 Importance level of each question using Random Forest, expressed relative to the maximum 77
5.5 Predictions when asking 1 question 77
5.6 Predictions when asking 2 questions 78
5.7 Predictions when asking 3 questions 78
5.8 Prediction quality as a function of number of questions asked 79
5.9 Output of the model with 10 questions 80
5.10 Back testing of model using 10 questions 81
5.11 result of the model with 1 question 82
5.12 Telemarketing 83
5.13 Back testing of calculated probability for manual underwriting (0 = accept, 1 = reject) 86

List of Tables

4.1 Chi-square test of Independence for categorical independent variables 48
4.2 Univariable analysis for continuous independent variables 49
4.3 Results of the `vif` function 52
4.4 Results of the `vif` function after excluding correlated variables 54
4.5 Hosmer-Lemeshow test 56

Appendices

Appendix A

Result of the glm function in R

```
Call:
glm(formula = DECISIONCA ~ PRODUCT + DEBITS + OCCUPATION + GENDERCA +
    AGE + SMOKERCA + HEIGHT + WEIGHT + BMI + PRODUCTCOVERAGEAMOUNT,
    family = "binomial", data = MPS)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8452  -0.4355  -0.2841  -0.1317   3.1738

Coefficients:
                                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                                     -4.937e+01  2.032e+03  -0.024  0.98062
PRODUCTLife                                     -4.544e-01  2.537e-01  -1.791  0.07331 .
PRODUCTLTC                                      -8.698e-01  5.000e-01  -1.740  0.08190 .
DEBITS                                           1.996e-02  1.366e-02   1.462  0.14386
OCCUPATIONAGENTE DI COMMERCIO                   -3.038e-01  2.431e+03   0.000  0.99990
OCCUPATIONAGENTE DI POLIZIA AMMINISTRATI        -1.892e+00  6.832e+03   0.000  0.99978
OCCUPATIONAGENTE DI POLIZIA OPERATIVO            7.372e-01  5.040e+03   0.000  0.99988
OCCUPATIONALBERGATORE/RISTORATORE                1.609e+01  2.032e+03   0.008  0.99368
OCCUPATIONALTRE CATEGORIE - ALTRI                1.421e+01  2.032e+03   0.007  0.99442
OCCUPATIONARCHITETTO                             1.556e+01  2.032e+03   0.008  0.99389
OCCUPATIONARTIGIANO                              1.563e+01  2.032e+03   0.008  0.99386
OCCUPATIONARTISTA/AUTORE E ASSIMILATI            1.653e+01  2.032e+03   0.008  0.99351
OCCUPATIONAUTISTA AUTOCARRI, TIR, BUS            1.568e+01  2.032e+03   0.008  0.99384
OCCUPATIONAVVOCATO                              -5.538e-01  2.749e+03   0.000  0.99984
OCCUPATIONBENESTANTE/REDDITIERO/PROPR.            1.553e+01  2.032e+03   0.008  0.99390
OCCUPATIONCARABINIERE                            3.721e-02  3.815e+03   0.000  0.99999
OCCUPATIONCARABINIERE OPERATIVO                 -5.847e-01  3.660e+03   0.000  0.99987
OCCUPATIONCASALINGA                              1.572e+01  2.032e+03   0.008  0.99383
OCCUPATIONCOLTIVATORE DIRETTO                   -2.851e-01  2.426e+03   0.000  0.99991
OCCUPATIONCOMMERCIALISTA/FISCALISTA             -4.988e-01  2.785e+03   0.000  0.99986
OCCUPATIONCOMMERCIANTE DETTAGLIO                 1.493e+01  2.032e+03   0.007  0.99414
OCCUPATIONCOMMERCIANTE INGROSSO                 -1.519e-01  2.874e+03   0.000  0.99996
```

OCCUPATIONCONIUGE DIPENDENTE	9.611e-01	4.244e+03	0.000	0.99982	
OCCUPATIONCONIUGE NON A CARICO	-1.455e+00	5.040e+03	0.000	0.99977	
OCCUPATIONDIRIGENTE/FUNZIONARIO IND/COM	-9.447e-01	2.908e+03	0.000	0.99974	
OCCUPATIONDIRIGENTE/FUNZIONARIO STATO	-1.178e+00	5.021e+03	0.000	0.99981	
OCCUPATIONDITTA INDIVIDUALE	3.762e+01	6.832e+03	0.006	0.99561	
OCCUPATIONELETTRICISTA ALTA TENSIONE	-4.869e-02	2.817e+03	0.000	0.99999	
OCCUPATIONFAMILIARI DIP MPS	1.621e+00	6.832e+03	0.000	0.99981	
OCCUPATIONFARMACISTA	-9.300e-01	4.247e+03	0.000	0.99983	
OCCUPATIONFIGLI DIP. NON CONVIVENTMPS	2.026e-01	5.040e+03	0.000	0.99997	
OCCUPATIONFORZE ARMATE NO MISSIONE	-1.689e-01	3.188e+03	0.000	0.99996	
OCCUPATIONFORZE ARMATE/AGENTE DI STATO	-2.400e-01	3.514e+03	0.000	0.99995	
OCCUPATIONGEOMETRA/DISEGNATORE	4.566e-01	5.040e+03	0.000	0.99993	
OCCUPATIONGIORNALISTA	-2.089e+00	5.025e+03	0.000	0.99967	
OCCUPATIONGUARDIA GIURATA O NOTTURNA	5.737e-01	6.832e+03	0.000	0.99993	
OCCUPATIONIMPIEGATO IND/COMM/SERVIZI	1.556e+01	2.032e+03	0.008	0.99389	
OCCUPATIONIMPIEGATO STATO/PARASTATO	1.539e+01	2.032e+03	0.008	0.99396	
OCCUPATIONIMPREND./ARTIG./COMMERC. ALTRI	1.523e+01	2.032e+03	0.007	0.99402	
OCCUPATIONIMPRENDITORE AGRICOLO	1.595e+01	2.032e+03	0.008	0.99374	
OCCUPATIONIMPRENDITORE INDUSTRIALE	1.462e+01	2.032e+03	0.007	0.99426	
OCCUPATIONIMPRENDITORE SERVIZI	1.482e+01	2.032e+03	0.007	0.99418	
OCCUPATIONINGEGNERE	1.576e+01	2.032e+03	0.008	0.99381	
OCCUPATIONINSEGNANTE	1.460e+01	2.032e+03	0.007	0.99427	
OCCUPATIONLAVORATORE EDILE	1.635e+01	2.032e+03	0.008	0.99358	
OCCUPATIONLAVORATORI DIPENDENTI - ALTRI	1.529e+01	2.032e+03	0.008	0.99400	
OCCUPATIONLIBERI PROFESSIONISTI ALTRI	1.551e+01	2.032e+03	0.008	0.99391	
OCCUPATIONMAGISTRATO	3.503e+01	5.040e+03	0.007	0.99445	
OCCUPATIONMARITTIMO O PESCATORE DI MARE	-5.843e-01	3.493e+03	0.000	0.99987	
OCCUPATIONMEDICO	1.629e+01	2.032e+03	0.008	0.99361	
OCCUPATIONMEDICO/DENTISTA	1.673e+01	2.032e+03	0.008	0.99343	
OCCUPATIONNOTAIO	-2.726e+00	5.040e+03	-0.001	0.99957	
OCCUPATIONOPERAIO IND/COMM/SERVIZI	1.567e+01	2.032e+03	0.008	0.99385	
OCCUPATIONOPERAIO STATO/PARASTATO	-4.794e-01	3.745e+03	0.000	0.99990	
OCCUPATIONPARACADUTISMO MILITARE NO GARE	3.797e+01	5.040e+03	0.008	0.99399	
OCCUPATIONPARAMEDICO	-1.792e+00	2.926e+03	-0.001	0.99951	
OCCUPATIONPENSIONATA EX DIPENDENTE DEL SETTORE PRIVATO	1.776e+01	2.032e+03	0.009	0.99303	
OCCUPATIONPENSIONATA EX LAVORATRICE AUTONOMA	3.416e+01	5.040e+03	0.007	0.99459	
OCCUPATIONPENSIONATI	1.845e+01	2.032e+03	0.009	0.99276	
OCCUPATIONPENSIONATO	3.673e+01	4.059e+03	0.009	0.99278	
OCCUPATIONPENSIONATO SOCIALE	-1.214e-01	5.025e+03	0.000	0.99998	
OCCUPATIONPERS.LE VOLO CIVILE	-1.719e+00	4.268e+03	0.000	0.99968	
OCCUPATIONRADIOLOGO, TEC. DI RADIOLOGIA	-3.253e-01	5.025e+03	0.000	0.99995	
OCCUPATIONRUSPISTA, TRATTORISTA	-7.002e-01	3.811e+03	0.000	0.99985	
OCCUPATIONSTUDENTE	4.692e-02	2.869e+03	0.000	0.99999	
OCCUPATIONTORNITORE,FRESATORE,FONDITORE	-1.256e+00	4.244e+03	0.000	0.99976	
OCCUPATIONVETERINARIO/AGRONOMO E ASS.	-9.669e-01	3.833e+03	0.000	0.99980	
GENDERCA1	-7.726e-01	3.097e-01	-2.495	0.01260	*
AGE	7.491e-02	1.170e-02	6.402	1.53e-10	***
SMOKERCA1	8.533e-01	2.754e-01	3.098	0.00195	**

```
HEIGHT                                                 4.876e-01  1.885e-01   2.587  0.00967 **
WEIGHT                                                -1.143e-01  4.218e-02  -2.711  0.00671 **
BMI                                                    5.734e-01  2.668e-01   2.149  0.03161 *
PRODUCTCOVERAGEAMOUNT                                  2.304e-06  3.084e-06   0.747  0.45511
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 930.94  on 1501  degrees of freedom
Residual deviance: 699.39  on 1428  degrees of freedom
AIC: 847.39

Number of Fisher Scoring iterations: 17
```

Bibliography

[1] Kroll Alice et Testa Ernest : Predictive modeling for life insurance seminar. *Society of Actuaries*, 2010.

[2] Leo Breiman : Random forests. *Machine learning*, 45(1):5–32, 2001.

[3] Vincent Calcagno, Claire de Mazancourt *et al.* : glmulti: an r package for easy automated model selection with (generalized) linear models. *Journal of Statistical Software*, 34(12):1–29, 2010.

[4] Chao Chen, Andy Liaw et Leo Breiman : Using random forest to learn imbalanced data. *University of California, Berkeley*, 2004.

[5] Mark S Dion et FLMI FALU : Predictive modeling: A life underwriter's primer. *On the Risk*, 27(2):36–43, 2011.

[6] Jerome Friedman, Trevor Hastie et Robert Tibshirani : *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

[7] Isabelle Guyon et André Elisseeff : An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[8] David W Hosmer Jr, Stanley Lemeshow et Rodney X Sturdivant : *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[9] Gary King et Langche Zeng : Logistic regression in rare events data. *Political analysis*, 9(2):137–163, 2001.

[10] Christopher Manning : Logistic regression (with r), 2007.

[11] Peter McCullagh et John A Nelder : *Generalized linear models*, volume 37. CRC press, 1989.

[12] Ricco Rakotomalala : Pratique de la régression logistique. *Régression Logistique Binaire et Polytomique, Université Lumière Lyon*, 2, 2011.